Supporting Exploratory Video Retrieval Tasks with Grouping and Recommendation

Martin Halvey[1], David Vallet[2,4], David Hannah[3], Joemon M. Jose[3]

[1]*University of Strathclyde, Glasgow, G1 1XQ, Scotland, UK*
[2]*Universidad Autonoma De Madrid, Madrid, 28049, Spain*
[3]*University of Glasgow, Glasgow, G12 8QQ, Scotland, UK*
[4]*National ICT Australia, 2217 Eveleigh, Sydney, Australia*

Abstract

In this paper, we present ViGOR (Video Grouping, Organisation and Recommendation), an exploratory video retrieval system. Exploratory video retrieval tasks are hampered by the lack of semantics associated to video and the overwhelming amount of video items stored in these types of collections (e.g. YouTube, MSN video, etc.). In order to help facilitate these exploratory video search tasks we present a system that utilises two complementary approaches: the first a new search paradigm that allows the semantic grouping of videos and the second the exploitation of past usage history in order to provide video recommendations. We present two types of recommendation techniques adapted to the grouping search paradigm: the first is a global recommendation, which couples the multi-faceted nature of explorative video retrieval tasks with the current user need of information in order to provide recommendations, and second is a local recommendation, which exploits the organisational features of ViGOR in order to provide more localised recommendations based on a specific aspect of the user task. Two user evaluations were carried out in order to 1) validate the new search paradigm provided by ViGOR, characterised by the grouping functionalities and 2) evaluate the usefulness of the proposed recommendation approaches when integrated into ViGOR. The results of our evaluations show 1) that the grouping, organisational and recommendation functionalities can result in an improvement in the users' search performance without adversely impacting their perceptions of the system and 2) that both recommendation approaches are relevant to the users at different stages of their search, showing the importance of using multi-faceted recommendations for video retrieval systems and also illustrating the many uses of collaborative recommendations for exploratory video search tasks.

Keywords: Video, search, user interface, collaborative, implicit, feedback, recommender, user studies.

1. Introduction

As a result of the improving capabilities and the declining prices of current hardware systems, there are increasing possibilities to store and manipulate videos in a digital format. 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. However, the systems that currently exist to organise and retrieve these videos are insufficient for dealing with such large and increasing volumes of video. In particular, there is a growing need to develop tools and techniques to assist users in the complex task of searching for video; this is particularly true online with the increasing growth of online video search systems.

Current state of the art video retrieval systems rely on textual descriptions or methods that use low-level features (e.g., visual features such as colour, shape, or texture; audio features such as the Fourier transform or pitch; and additional features such as automatic speech recognition (ASR) or optical character recognition (OCR)) to find relevant videos within a large collection. Neither of these methods is sufficient to overcome the problems associated with video search. On the one hand, query by text relies on the availability of sufficient textual descriptions of the video and its content, resulting in a heavy system dependence on users providing relevant text descriptions and annotations. The main drawback of this approach is that often users can have very different perceptions about the same video and annotate that video differently [13], which makes it difficult for different 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 [14], resulting in a lack of available textual annotations. On the other hand, 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 [29], provides difficulties for using these low-level features. Consequently, while these low-level features are used in some state of the art systems, most online video retrieval systems (e.g. YouTube[1] or Blinkx[2]) rely only on query by text.

In order to alleviate some of these problems associated with video search we have developed ViGOR, a video retrieval system that allows users to create semantic groups of results to help conceptualise and organise their results for complex video search tasks. This interactive grouping is a flexible means for a user to illustrate their multi-faceted information needs. Multi-faceted information needs mean that the task that the user is conducting can be considered to be multiple specific tasks; it could also be considered that multi-faceted search tasks/information needs can have multiple solutions. Specific information needs can be related to short term information needs as the user is focused on one particular aspect of their search task. The grouping facilities also allow the user to focus on one particular aspect of a global task as they can focus on specific (or short-term) information needs while still solving the overall multi-faceted (or long-term) information need as embodied by their search task. We believe that the semantic gap is narrowed by this abstraction to high-level semantic groupings reflecting an individual's task-specific mental model of the data and a more flexible user interaction with the video collection, thus the user is focused more on interaction with the data and less on the mechanics of their search. We also believe that the use of this system can result in a number of desirable outcomes for users: improved user performance in terms of task completion and task exploration, and increased user satisfaction with their search and their search results.

In addition, the interactions available in ViGOR make it an ideal system with which to integrate some recommendation techniques. We believe that many of the problems associated with searching large collections of video can be alleviated through the use of recommendation techniques. Recommendation techniques can offer a work around for the problems associated with the semantic gap and the unreliability of textual descriptions, as they utilise additional information about user interaction that is already available in many systems. However, it is also imperative that the recommendations relate to as many aspects of a user task as possible so as to ensure that the recommendations present the user with a diverse set of results that encompass as many interpretations of the user actions as possible. To that end, we have developed a recommendation approach that utilises the implicit actions involved in previous user searches to create a predictive model that can provide multi-faceted and diverse recommendations to assist users in completing their difficult search tasks. Our recommendations encompass many interpretations of user actions and numerous videos that users may not have seen using normal query methods. Providing these recommendations is not trivial, as due to the complex and difficult search process for video, implicit feedback from video search is quite noisy [29]. However, we believe that this problem can be overcome by utilising collaborative recommendation techniques. In particular we believe that our approach of modelling many aspects of user needs via implicit user interactions can result in improved user performance in terms of task completion and reduce the user effort involved in finding relevant videos.

Before proceeding, it should be noted ViGOR has been developed as an interface that can sit on top of any video retrieval system. The recommendation and grouping facilities provided as part of the ViGOR system can be created based on the log files stored by almost any system, website etc. As such while the recommendation and interface are coupled in this system, in some ways they can be viewed as two distinct parts that can be applied to any video retrieval system, however in this case we are attempting to leverage the benefit of both. Thus as well as solving problems surrounding exploratory video search tasks, we have developed a scalable solution which can be deployed on top of almost any existing video retrieval system anywhere (this has been demonstrated by conducting evaluations using ViGOR in conjunction with YouTube, the largest online video storage and retrieval system (see Section 6 and Section 7)).

In order to test and validate the potential benefits of ViGOR in assisting with video search, we conducted two user studies, in which we tested two systems. The first ViGOR without recommendations, and the second a system based on ViGOR that provides recommendations based on a model of implicit actions. These systems were evaluated to determine

[1] http://www.youtube.com/

[2] http://www.blinkx.com/

whether any benefit to the users is achieved. The remainder of this paper is organised as follows: in the following section we will provide a rational for this work. Section 3 will describe the two systems that were used in our study. Subsequently, in section 4 we will describe our approach for using implicit feedback to provide multi-faceted recommendations. In section 5 we will describe our experimental methodology including our hypotheses, which is followed by the results of our experiments, presented in Section 6 and 7. In Section 8 we will provide a discussion of our work and section 9 will provide some final conclusion and a discussion of future directions for this work.

2. Related Work

2.1 Interactive video retrieval

Interactive video retrieval consists of users formulating queries and carrying out video searches, and then reformulating those queries and thus the current results based on previously retrieved results. As video is in essence multimodal i.e. it consists of a variety of content types, there are a variety of methods that can be used to query a video retrieval system. One approach commonly adopted is to use the low-level features that are available in images and videos, i.e. colour, texture, shape etc. for retrieval. This is often utilised in a query by example approach; using query by example users provide examples via sample images or video clips in order to retrieve comparable images or video clips. While this approach seems both intuitive and reasonable, there are also a number of limitations. Query by example requires the extraction, representation and storage of a set of low level features from all of the videos in the collection, this presents issues related to efficiency. Also the semantic gap [29], provide difficulties. The semantic gap essentially is the difference between the high level concepts users associate with video and the low-level data representation of videos used for visual features. Overcoming the problems associated with the semantic gap is one of the largest and most challenging research issues in multimedia information retrieval. As a consequence of the semantic gap and in an attempt to overcome the limitations it causes, the multimedia search research community has investigated search by concept. The basic idea behind search by concept is that more semantic concepts such as "vehicle" or "person" can be used to aid retrieval instead of relying on just low level features; an example of this is the Large Scale Ontology for Multimedia (LSCOM) [25]. However as it is still in the early stages of research query by concept also has a number of issues that hinder its widespread use, the biggest problem is that it requires a large number of concepts to be represented and as a result to date it has not been deployed on a large scale for general usage.

The most common method of searching for video does not use low level features or concepts, but is rather query by text. Query by text is the approach used in many online large scale video retrieval systems, e.g. YouTube, MSN Video, or Blinkx, and is also the most popular query method at TRECVID [5]. Query by text is simple and users are familiar with this paradigm from other types of search. In addition, query by text does not require a representation of concepts or features associated with a video. However, it does require that meaningful textual descriptions of the content of the video are available and this is not always the case. 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 [19] concludes that the availability of these additional resources varies for different systems. Where these descriptions are available there may be some reliability issues, this could be due to a number of factors, e.g. limitations in automatic speech recognition, language differences etc. Most contemporary online video search systems rely mostly on annotations provided by users to provide video descriptions. However, as was stated previously, this further complicates the retrieval process, either because of misconception surrounding annotations [13] or users' reluctance to provide annotations [14].

While the methods outlined above have some limitations and problems, they have been used together in a number of systems, e.g. Informedia [17] and MediaMill [30]. These systems have been some of the best performing systems at past TRECVID interactive search evaluations. However, these best performing results are quite often for "expert" users, who establish an idealistic upper bound of performance of video search users [6]. In addition, a combination of these approaches requires a vast amount of metadata to be extracted and stored for each individual video clip.

As has been outlined, there are a number of different ways in which a user can query a video retrieval system; including query by text, query by example and query by concept. However, each of these methods have had limited success in solving

the problems associated with video retrieval. To date none of these approaches has provided a complete and wholly adequate solution to providing the tools to facilitate video search [5]. Thus, in an attempt to overcome these limitations some innovative interfaces have been proposed for multimedia search, some of these interfaces are described in the following section.

2.2 Interactive Multimedia Retrieval Interfaces

As has been described above there are a number of problems which hinder effective multimedia retrieval. Some of these problems can be overcome by providing effective interfaces for interactive multimedia retrieval. In this sub-section we outline some innovative and interesting interfaces for interactive multimedia retrieval. PicturePiper [10] provides a mechanism for allowing users access to images on the web related to a topic of interest. This system was developed to demonstrate a re-configurable pipeline architecture that is ideally suited for applications in which a user is interactively managing a stream of data. PicturePiper also contains a workspace for displaying search results; the distance between groups of images in the workspace illustrates the distance between the centroids of the groups of images as calculated using low level features. CueFlik [11] is a Web based image search application that allows users to create their own rules for ranking images based on their visual characteristics. Users can then re-rank possible search results according to these rules. In user evaluations it was found that users can quickly create effective rules for a number of diverse concepts. EGO [33] is a tool for the management of image collections. The main components of EGO are a workspace and a recommendation system. By providing these facilities, different types of requirements are catered for, enabling the user to both search and organise results effectively. The workspace serves as an organisational ground for the user to construct groupings of images. The recommendation system in EGO observes the user's actions, which enables EGO to adapt to their information requirements and to make suggestions of potentially relevant images based on a selected group of images. ImageGrouper [24] is another interface for digital image search and organisation. It is possible to search, annotate, and organise images by dragging and grouping images on the workspace, ImageGrouper allows users to use an entire group as a query by example and allows a user to annotate an entire group at once. However, no formal evaluation of ImageGrouper has taken place. The MediaGLOW system [12] presents an interactive workspace that allows users to organise photographs. Users can group photographs into stacks in the workspace; these stacks are then used to create neighbourhoods of similar photographs automatically. Campbell presents a novel image search and browsing system in the Ostensive Browser [4]. The main component of the interface is a workspace with objects on it and links between those objects. The user begins browsing at a starting image, around which candidate next images are displayed. A user selects a candidate which becomes the centre of focus; the next possible candidate images related to the current image are displayed. Browsing continues in this fashion. Candidate images for browsing are determined by an ostensive model, which encompasses a temporal profile of uncertainty. This is accomplished by the application of a particular class of discount function with respect to the age of the evidence, thus more recent user interactions are more relevant for determining next steps in comparison with older interactions.

While the systems mentioned above have made a number of advances in relation to image search, there are a number of important differences that make video search much more difficult than image search. The first main difference is the multimodal nature of video, encompassing images, text, audio and a temporal factor etc. While text and visual features may be used to aid or hamper image search, these are only two of the many modalities involved in video search. Secondly, video is a much more interactive medium in comparison with still images. Interactive video retrieval systems have to make an additional effort to aid the user in deciding whether the selected videos are relevant or not for their tasks, whereas for image retrieval systems the user can easily and quickly discern relevant and irrelevant results. The result of this is that interaction and usage information from interactive video retrieval systems is far noisier than the usage information on image retrieval systems. For instance, on average, 75% of the user results that the user interacted with on the image retrieval system developed by Craswell and Szummer were relevant [7], whereas only 7-9% of search results that the user interacted with were relevant for a similar interactive video retrieval system [21]. As a result of this, the goals of many interactive video retrieval systems are to lower the effort for the user to explore the complex information space and also to assist the user in deciding if a result is relevant to their information need. To that end a number of video retrieval systems have been developed

to aid user interaction. However, it should be noted that the ViGOR interface does not do anything beyond what any other video retrieval system that is mentioned here does to deal with the multimodal nature of video or indeed interactivity with the actual videos themselves. The focus the grouping in the interface is to overcome some of the problems caused by the semantic gap. However, our recommendation approach does take into account some of the actions that users can perform with video, e.g. playing the video, thus in this way some of the interactivity that can take place with video is taken into account.

The ForkBrowser [26] embeds multiple search methods into a single interface for browsing. The multiple search methods are presented to the user in the form of threads. These threads are ranked lists of shots based on one of the search methods implemented in the interface. The threads are visualised in the shape of a fork. The shot at the top of the stem of the fork is the video that the user is currently viewing, with the tines representing the different threads. The ExtremeBrowser [18] aims to maximise the human capability for judging visual material quickly, while at the same time applying active learning techniques using the user selected videos. Videos are presented to the user via a method called rapid serial visual presentation which allows the user to make fast judgements about high numbers of videos. The feedback from the user is used in an active learning loop, which is used to rank the remaining results that the user will review. The FacetBrowser [35] is a video search interface that supports the creation of multiple search "facets", to aid users carrying out complex video search tasks involving multiple concepts. Each facet represents a different aspect of the video search task: an assumption of this work is that search facets are best represented by sub-searches. These facets can be organised into stories by users, facilitating the creation of sequences of related searches and material which together can be used to satisfy a work task. The interface allows more than one search to be executed and viewed simultaneously, and importantly, allows material to be reorganised between the facets, acknowledging the inter-relatedness which can often occur between search facets. The goal of these systems is common: to overcome the semantic gap problem of video content. Another way of overcoming the semantic gap is through the use of recommendation techniques, of particular interest to the work outlined in this paper are techniques that use past usage history, a number of such approaches are outlined in the next section.

2.3 Usage-based Recommendation approaches in multimedia information retrieval

The usage history from a community of previous users can be an important source of information in order to improve the performance of Information Retrieval (IR) and Multimedia IR (MIR) systems; whenever a user enters a query, the system can exploit the behaviour of previous users that were performing a similar task [1] [7] [36] [21]. For instance, Bauer and Leake built up a task representation based on the user's sequence of accessed documents [1]. This task representation was used by an information agent, which proactively suggested documents to the user. One particular approach of past usage information exploitation is the use of click through data [7] [8] [32] [36] [21]. Click through data is limited to the query that the user executed into the system, the returned (multimedia) documents, and the subsequent documents that the user opened to view. Craswell and Szummer represent the click through data of an image retrieval system as a graph, where queries and documents are the nodes and the links are the click through data [7]. White et al. introduced the concept of query and search session trails, where the interaction between the user and the retrieval system is seen as a path that leads from the first query to the last document of the query session or the search session (i.e. multiple queries) [36]. They argue that the last document of these trails is more likely to be relevant for the user. In a more recent study, White and Huang [37] concluded that exploiting the full query trail instead of the last accessed document resulted in recommendation with higher values of topic coverage and diversity, without necessarily harming the expected relevance and utility of the documents belonging to the search trail. Their hypothesis coincides with ours: when users perform complex search tasks, usually they find diverse results during the search exploration, which cover different aspects of the initial query, as the initial information need of the user evolves. In their work, White and Huang propose recommending a related query trail to users when issuing a query. In complementary work, Singla et al. [27] proposed a number of techniques for efficient query trail finding and recommendation, given an initial query and the first selection of a search result. They measure the effectiveness of the proposed approaches based on different metrics, such as coverage, utility, diversity, and relevance. The main difference between Singla et al.'s and our recommendation approach is that we propose to recommend individual results that can belong

to different search trails, instead of recommending a query search trail related to the user's search. Our approach also exploits session search trails that include multiple queries, instead of a single query trail. With this we expect to obtain even more diverse recommendations, as shifts in user needs are usually not restricted to a single query.

Hopfgartner et al. [21] expanded the above work on click through data by taking into consideration an MIR system, which includes other types of actions such as playing a video for a given duration or navigating through a video. Their goal is to exploit the community based feedback mined from the implicit interactions of previous users of their video retrieval system to aid users in their search tasks [21]. Some of this work has been used as a basis for the recommendation techniques that we have implemented in our own video retrieval system. Similar to White et al.'s [36] notion of search trails; we represent user interactions as a linked sequence of interacted documents and input queries during a search session. Following Craswell and Szummer [7] and Hopfgartner et al. [21], we adopt a graph-based approach, as it facilitates the representation of interaction sequences.

While most of the authors limit the graph to click through data, we propose to integrate other sources of implicit relevancy into the representation, following some of the work of Hopfgartner et al. [21]. As will be shown in the following section, ViGOR allows the collection of more and richer interactions, related to the interaction of users with ViGOR's grouping functionalities. In addition, our recommendation approaches couple the multi-faceted and ambiguous nature of explorative tasks, where a user can typically interact with different aspects of a retrieval tasks within the same session. This is achieved through the representation of soft-links, introduced in Section 4.2. The ViGOR system for searching and recommendation is presented in the next section.

3. System Description

3.1 ViGOR: A Video Grouping and Organization Interface for Video Retrieval

The main goal of ViGOR is to provide grouping functionalities for interactive video retrieval tasks. ViGOR (see Figure 1) comprises of a search panel (A), results display area (B), workspace (C) and playback panel (D). These facilities enable the user to both search and organise results effectively. The users enter a text based query in the search panel to begin their search session. The result panel is where users can view the search results (a). Additional information about each video shot can be easily retrieved. Placing the mouse cursor over a video keyframe for longer than 3 seconds will result in any text associated with that video being displayed to the user (we will hence forth refer to this action as tooltip) (e). If a user clicks on the play button the highlighted video shot will play in the playback panel. Users can play, pause, stop and navigate through the video as they can on a normal media player.

Similar to the MediaGLOW [12], PicturePiper [10], ImageGrouper [24] and EGO [33] systems, one of the main components of ViGOR is the provision of a workspace (C). In PicturePiper [10] the workspace is merely used to visualise the difference in search results. MediaGLOW, ImageGrouper and EGO use the workspace to group images only. However, as has been discussed (see Section 2.2 for full details) the problems associated with video and image search are very different and we believe that the approach of using groups in a workspace is an extremely useful solution for video search, as we will demonstrate in this paper (see section 5.1 for some additional details on our hypothesis). In addition, ImageGrouper and EGO were evaluated on small controlled collections of images, whereas we are testing ViGOR in an online scenario for a huge video collection i.e. YouTube. In ViGOR, the workspace serves as an organisation ground for the user to construct groupings of videos, these groups can relate to different aspects of the users search task. This facility allows the user to express different aspects and organise the results of their search in whatever way they want. The FacetBrowser [35] also allowed users to organise video search results, however the focus of the FacetBrowser was to merely allow the user to view multiple search threads and re-organise those threads. ViGOR is much more interactive and allows the user to store and mix results from previous searches to inform future searches, while at the same time allowing the user to carry out searches independent of these groups if the user wishes. Groups can be created by clicking on the create group button. Users must then select a textual label for the group and can potentially add any number of annotations to the group, but each group must have at least

one annotation. Drag-and-drop techniques allow the user to drag videos into a group or reposition the group in the workspace. It should be noted that any video can belong to multiple groups simultaneously.

The workspace is designed as a potentially infinite space to accommodate a large number of groups. Each group can also be used as a starting point for further search queries. Users can select particular videos in the group's panel (b) and can choose to view an expansion of the group that contains similar videos based a number of different features (d). We will call this functionality local expansion. As the ViGOR system uses YouTube as a backend, the features available to perform a local expansion of the group are mainly standard YouTube features. The interface offers three expansion options (c): 1) text expansion, which is the result of a new search using text extracted from the selected videos; 2) related videos; and 3) videos from the same user. All of the videos returned by these expansion options are retrieved using the YouTube API. Some of the interface components allow users of the system to provide implicit feedback, which is then used to provide recommendations to future users. Implicit feedback is given by users adding a video to a group panel (b), playing a video (D), highlighting a video using the tooltip (e), or submitting a search query (A).

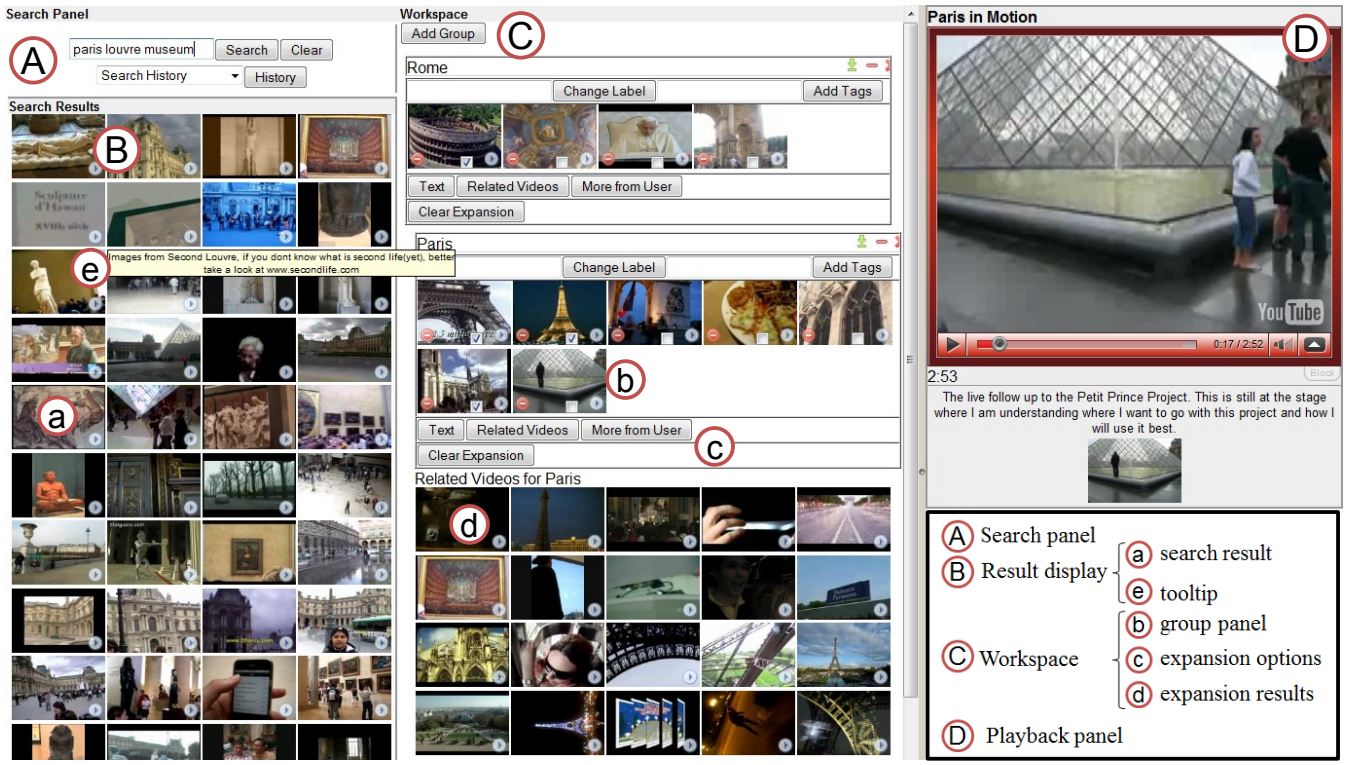

Figure 1. ViGOR interface.

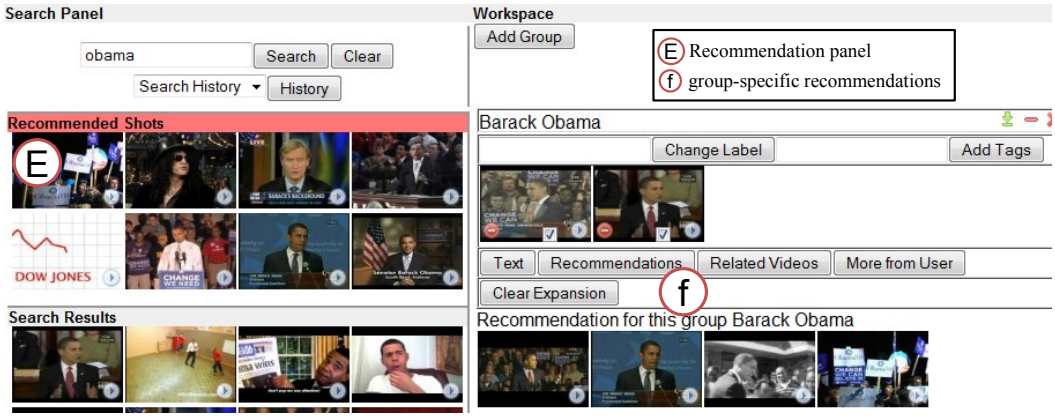

Figure 2. Changes for the ViGOR interface with recommendations.

3.2 Integrating Multi-faceted Recommendations into ViGOR

The extended ViGOR with recommendations system is designed as an extension of the ViGOR system (See Figure 2 for a detail on the changes of the interface). No functionality is removed, ViGOR with recommendations still allows users to organise their search in different groups in the workspace and allows the execution of text queries along with the local expansion functionalities (i.e. related videos, videos from same user, text expansion). In the extended interface with recommendations, users also have two options to receive recommendations. The users are presented with recommendations of video shots that might match their search criteria based on their interactions (E); these are global recommendations that incorporate the entire search session (see Section 4.4). These global recommendations are updated whenever the users plays a video, issues a new text query or moves a video to a group panel. One new option for a local expansion is added: the local recommendation (f), with which users may also retrieve recommended videos within a group. These recommendations are localised to each group and are based on the interactions of previous users with videos that the current user has selected (see Section 4.5). The following section will provide further details about the recommendations that are provided by the ViGOR system.

4. A Multi-faceted Graph Based Recommendation Approach

In this section we introduce the multi-faceted recommendation approach integrated into the ViGOR system. We present two recommendation techniques. The first is a global recommendation technique, which is an extension of a previous recommendation approach [21] that takes into consideration the new interactions provided by ViGOR, and incorporates the concept of soft-links for multi-faceted and diverse recommendations. The second approach is a novel local recommendation technique which has been specifically defined to take advantage of the grouping facilities provided by ViGOR. The goal of these recommendation approaches is twofold: 1) to exploit the organisational functionalities provided by ViGOR as a new source of implicit information; and 2) to take into consideration the ambiguous and multi-faceted nature of an exploratory video search, through the use of soft links. To the best of our knowledge, these points have not previously been addressed.

4.1 Weighted Graph User Interaction Representation

We follow a graph-based approach for the representation of past user interactions. In this approach, a user session s is represented as a set of queries Q_s, which were input by the user, the set of multimedia documents (in the case of this work videos, however multimedia documents can refer to any type of document) D_s the user accessed during the session, and a set of groups or aspects G_s the user created during the search session. Queries, documents and groups are thus the nodes $N_s = Q_s \cup D_s \cup G_s$ of our graph representation $G_s = (N_s, W_s)$. The arcs of this graph representation, W_s, are of the form $W_s = (n_i, n_j, u, w_s)$ and indicate that at least one action led the user u from the node n_i to n_j. Note that the only action that can lead to a group node $g \in G_s$ is the action of moving (i.e. assigning) a document node to the group. 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 estimated by means of considering the implicit evidence given by each type of action by users with that node, following a previously developed implicit model [20]. Users' interactions can be represented within the same weighted graph, as they share the same node representation (e.g. in the investigated interface, actions such as playing, tooltipping or selecting as relevant a video were represented into the graph). Query nodes are simply identified by the query terms, documents are identified by their URL and group nodes are identified by the containing documents. Groups can be labelled by users, but this information is not used to identify groups within the system. This means that whenever two users issue the same query or interact with the same document, the sessions will be connected in the final graph representation, indicating there is some sort of relation between both users' sessions. Note that this graph is not necessarily fully connected, as there could be clusters of sessions with no nodes in common. In summary, one important property of this graph representation is that it allows the

agglomeration of all past user interaction in a single aggregated pool of past usage information, which we will henceforth refer to as the *implicit pool*. See Section 4.3 for a further explanation of this technique.

4.2 Soft Link Motivation

In recommendation algorithms based on a graph-based representation of past usage history [7] [36] [21], the nodes of the graph indicate user queries or documents in the search collection, whereas each link indicates a transition of the user from one content to another. For instance, in the example graph shown in Figure 3, the vertical path represents a user that input the query 'Paris' in the search system and then accessed consecutively the search results represented by nodes n_1, n_2, and n_3. These graphs are later exploited in a number of ways in order to recommend results to the current user.

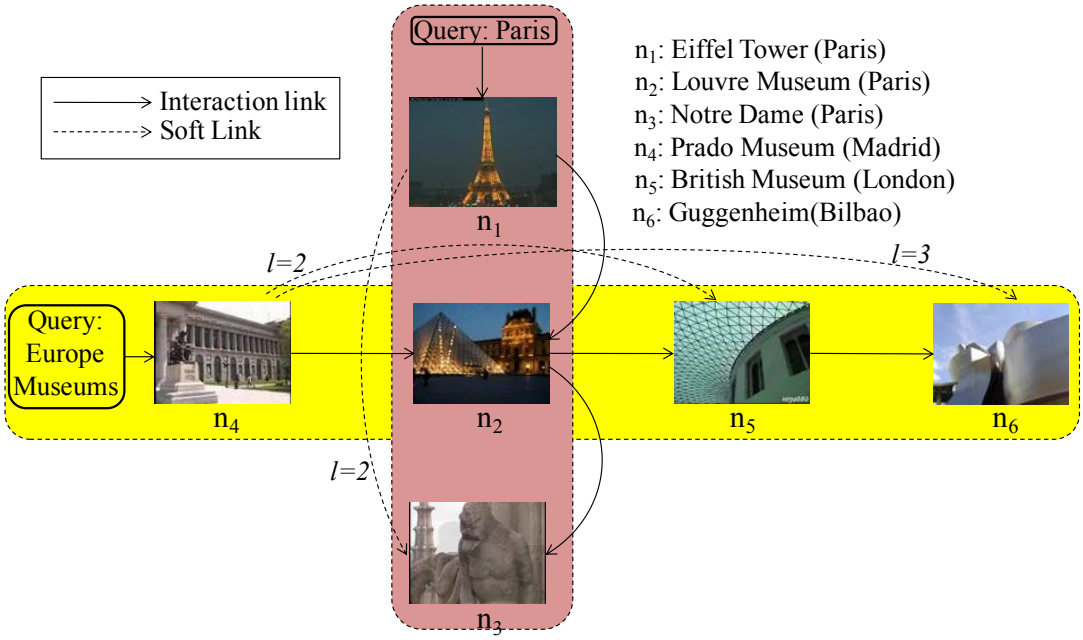

Figure 3. Example of user interaction and soft links in two related task.

However, many techniques do not take into account the potential ambiguous and/or multi-faceted nature of some tasks, and thus treat each possible aspect inspected by the user as if it was the same or indeed if each aspect is related to a single task. To illustrate this we follow the simplified example of a graph-based usage representation of the user actions over the system, depicted in Figure 3. It can be observed that there are two different aspects stored in this graph, the first one regarding the city of Paris, and the second regarding different European museums. These aspects could be part of a more complex search task, such as "find places to visit in Europe". These two aspects intersect in one node: 'Paris Louvre Museum'. Let us suppose that the current interactions sequence of a user is related to nodes n_4 and n_2 (e.g. because she has opened these items). Recommendation approaches that use the direct links [7] [21] will score nodes n_5 and n_3 the same, because they directly follow node n_2. However, we believe that it is more sensible to give a higher score to n_5 as it belongs to the same search aspect that the user seems to be following. White et al.'s recommendation approach [36][27], based on search trails, can also give some consideration to the actual aspect of the search session, but they rely on having information on the initial query input by users, while our approach only needs the interaction information regarding documents. In addition, as is shown in the experiments section (see Section 7.3), White et al.'s approach also benefits from this representation.

Motivated by the previous example, we propose the use of soft links in order to overcome the problems of aggregating all user interactions into one single graph-based representation. Soft links are arcs that create soft relations between nodes that belong to the same interaction sequence of user actions, indicating the level of distance separation (l in Figure 3). The purpose of soft links is to maintain the original subsequent interactions of a user even if his/her interaction data is aggregated

with other users' interaction information. In this way, we can take the current aspect of the user's interactions into consideration when making a recommendation.

Furthermore, as soft links help to disregard which aspects are parts of the trail of interactions, a recommendation approach that uses soft links might be able to diversify its results by aggregating recommendation results at different levels of soft links. For instance, in following the previous example, when using the original interaction links, i.e. a soft link level of $l = 1$, the recommendation approach would recommend node n_3, among others, from the "city of Paris" aspect. On the other hand, when using soft links of level $l = 2$, the recommendation algorithm will ignore the original links and, in this example, recommend n_5, which belongs to the "European Museums" aspect. Furthermore, as more levels of soft links are aggregated it is expected that more aspects will be found, as the recommendation algorithm advances further through the interaction information of past users, which could contain changes of aspect of their actual search, in case of their task being explorative and multi-faceted. We thus put forward that a proper aggregation of all the recommendations for all levels of soft links will provide recommendations that are not only relevant to the current task of the user, but that belong to different aspects related to the search tasks and thus are more diverse than only considering the direct interaction information.

In conclusion, we hypothesise that soft links enable further exploration in the usage information collected from past users, distinguishing and reaching other aspects related to their multi-faceted search. Thus a recommendation approach that uses soft links at different levels, instead of the original usage information, will be able to provide more diverse recommendations. In the following section we provide a formal definition of soft links and in Section 7.3 we will perform a user-centred experiment in order to validate our hypothesis.

4.3 Soft Link Weighted Graph

We hence complete the weighted graph outlined above by the use of *soft links*. Soft links are modelled as special action arcs $W_s(l)$, represented in the form $\left(n_i, n_j, u, w_s(l)\right)$ where in the case there is a path $p = n_i \rightsquigarrow n_{j-1} \rightarrow n_j$ from n_i to n_j in the session graph. $n_{j-1} \rightarrow n_j$ means that n_{j-1} is adjacent to n_j, and the same notation is used as shorthand to define *p* as any path between n_i and n_j, taking into consideration the link directionality. The action's weight is obtained from the action arc $(n_{j-1}, n_{j,}, u, w_s)$ and l is equal to *length*(*p*), which is counted as the number of links in path *p*. A soft link of level l will thus connect each node with the node at a distance l of the same user's search interaction trail. Figure 3 shows a simplified example of a soft link weighted graph. Note that $W_s(1)$ is equal to W_s and thus we can represent the session graph as $G_s = (N_s, W_s(l))$ where $l \in \{1,2, \dots, L\}$ varies from 1 to the maximum level of soft links considered, denoted as L.

Finally, all the session graphs are aggregated into a single graph $G = (N, W(l))$, where $N = \bigcup_s N_s$ and $W(l) = \bigcup_s W_s(l)$, which constitutes a global pool of usage information that collects all the implicit relevance evidence of users from past sessions. This graph is the implicit pool of the past community of users. The nodes of the implicit pool are all the nodes involved in any past interaction $N = \bigcup_s N_s$, whereas the weighted links W(l) are of the form $\left(n_i, n_j, w(l)\right)$, where $n_i, n_j \in N$ and w combines the probabilities of all the session-based values for a specific soft link distance value of l. Note that $W(1)$ models the aggregation of past interaction sequence of users, whereas $W(l), l \in \{2, \dots, L\}$ models the aggregations of the soft links at different distance levels. As weight values are considered probabilities of relevance of the node n_j to the user, we opt for a simple aggregation of these probabilities, this is $w(l) = \frac{\sum_s w_s(l)}{\#w_s(l)|w_s(l)>0}$, which represents the average probability associated to n_j in any session that n_j was involved in. Each link represents the overall implicit (or explicit, if available) relevance that all users whom actions or soft links led from node n_i to n_j, gave to node n_j. The role of soft links is made clearer in this step. Without soft links, the user's original search interaction trail could be lost or blurred when interactions from other users are aggregated. With soft links, although links are agglomerated into a single implicit pool, the user interactions are aggregated using a number of different soft link levels, thus the user's original interaction trail is still represented and can thus be exploited by our global recommendation algorithm, introduced in Section 4.4.

Figure 4 shows a close up image of an implicit pool from our evaluations, in which we can see in more detail the interaction sequences and associated weights, the video nodes, the query nodes (marked with an asterisk), and the group panels nodes, which are related to the set of contained video nodes. In Figure 4 we can observe an example of two group panels, labelled as 'Barack Obama', which were created by two different users, and have some documents in common. This is the type of panel representations that can be exploited by the local recommendation approach, explained in Section 4.5

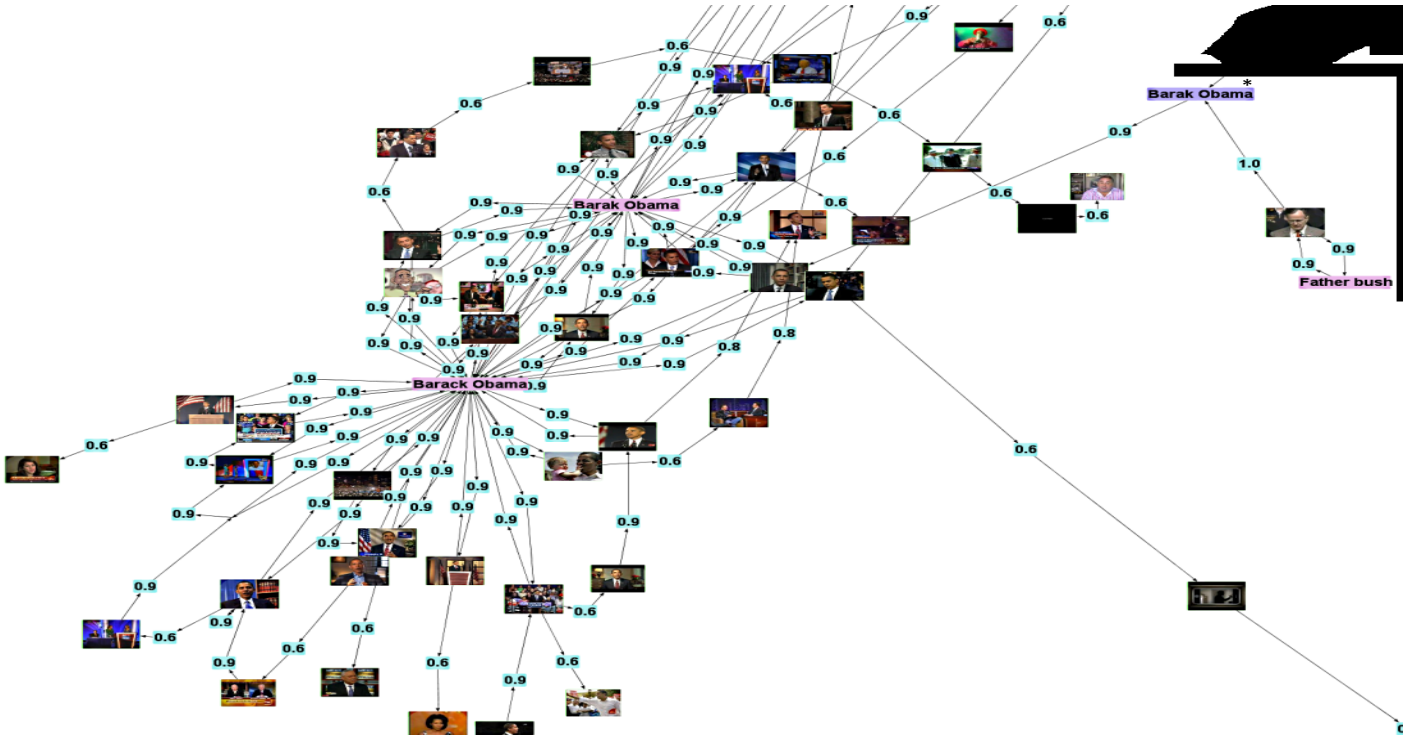

Figure 4. Detail of implicit pool. Text nodes with an asterisk indicate query nodes, whereas the rest of text nodes indicate group panel nodes. For the sake of clarity, soft links are not here represented.

4.4 Global Soft Link Recommendation

The global recommendation approach is based on the status of the current user session. As the user interacts with the system, a session graph $G_s = (N_s, W_s(l))$ is constructed, where in this case s is the current user's ongoing session. This graph is the input for the global recommendation algorithm presented next. This recommendation approach has two goals: 1) to exploit the implicit pool in order to retrieve similar nodes that were somewhat relevant to other users and 2) to exploit the soft links in a way that the user's outgoing aspect within their task is taken into consideration. This recommendation approach is defined in two steps. Firstly, the global recommendation is defined for each soft link level:

$$gr(n, N_s, l) = \sum_{\substack{n_i \in N_s \\ p = n_i \rightsquigarrow n_j \rightarrow n \\ length(p) < D_{MAX}}} lr'(n_i) \cdot \xi^{length(p)-1} \cdot w\left(l, n_j, n\right), \quad (1)$$

Where $n_i \rightsquigarrow n_j$ denotes the existence of a path from n_i to n_j in the graph, taking link directionality into consideration. $n_j \rightarrow n$ means that n is adjacent to n_j. $w\left(l, n_j, n\right)$ is the probability weight, given by the implicit pool, for a soft link distance of l, as this weight is exploited in order to rank n. $lr'(n_i) \in [0,1]$ is a weighting function that follows our previous implicit model [21] based on the relevance of this node to the outgoing user's session, obtained from the user's implicit feedback. *Length*(p) is counted as the number of links in path p, which must be less than a maximum length D_{MAX}, set to 5 in for our

evaluation. Finally, ξ is a length reduction factor, set to 0.8, which 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 may still be recommended. These values were tuned using development data such as the implicit information and relevance judgements collected from our initial user study with ViGOR. In a second step, the final recommendation score is computed as

$$gr(n, N_s) = \prod_{l=1}^{L} gr(n, N_s, l) \quad (2)$$

where L is the maximum level of soft link considered, set to 10 in our experiments. It was decided to only evaluate the recommendation of documents (videos in the case of our evaluation) to the user, so the ranking was filtered by $n_d = n \in \mathrm{D}$, i.e. by only nodes that belong to documents.

4.5 Local Group Recommendation

The local group recommendation focuses on a local expansion of a set of selected documents within a group, and tries to recommend more documents that could aid the user in expanding the aspect of the task represented by this group. This recommendation approach exploits the representation in the implicit pool of the different aspects created by previous users. In this case, the local group recommendation tries to find similar aspects that previous users could have created and then rank their related documents. The input of the local group recommendation is the set of documents $\mathrm{D_g}$ that the current user has selected within an aspect group $\mathrm{g} \in \mathrm{G}$. On the first step of this approach, related aspect groups from the implicit pool are searched, this is achieved by ranking the related groups panels $n_g \in \mathrm{G}$ with the global recommendation approach, using the set of selected documents as input. Thus, the related groups can be ranked, using Equation 2, as $gr(n_g, D_g)$, were the local relevance of the selected documents is set to 1, i.e. $lr'(d_g) = 1, d_g \in D_g$. We tuned the parameters for this approach in order to focus on the current aspect; this was done by setting the maximum soft link value L to 3. We also limited the expansion distance D_{MAX} to 3, in order to constrain the search to more highly related groups.

On the second step of this approach, the implicit pool is exploited in order to rank the top nodes related to the set of ranked group nodes n_g. Group nodes are connected to their contained documents by a special action, namely 'panel contains'. This allowed us to rank documents based on how many group panels contained them. Note that group panels are uniquely identified by the user who created them and their assigned title, so it is possible to have different group panels from different users which contain related and overlapping sets of documents (this phenomenon is illustrated in Figure 4). The ranking approach in this second step is to rank higher those documents that belong to more related aspects created from previous users. This ranking can be also implemented by tuning the global recommendation approach in the following way. The input of the ranking approach will be the ranked set of related groups, $n_g \in \mathrm{G}$. The local relevance of the input is the ranking given by the previous step: $lr'(n_g) = gr(n_g, D_g)$. Thus, the final ranking is obtained from $gr(n, n_g)$. No soft links are used in this ranking, and the expansion distance D_{MAX} is set to 1, as we only want to rank documents that are contained by at least one of the related group nodes. As with the previous approach we filter the recommended nodes to contain only documents $n_d = n \in \mathrm{D}$.

In the following sections, the outlined recommendation techniques will be evaluated and compared with the ViGOR system without recommendation, in order to assess the benefits and effects of adopting a multi-faceted recommendation approach. However, first we must assess the utility of the grouping paradigm that is used in ViGOR to assist user carrying out exploratory video search tasks.

5. Experimental Methodology

5.1 Hypothesis

In order to measure the effectiveness of our proposed approach we conducted two user-centred evaluations. The two user evaluations conducted were both between subject evaluations that involved users carrying out broad video search tasks on YouTube. This provided us with a large and dynamic data collection, and facilitated the analysis of ViGOR in an online situation. The first evaluation compared a baseline system, which mimicked YouTube's functionalities, with our own system ViGOR, without recommendations. We had three hypotheses to address in the first evaluation:

- *Hypothesis H1.1: Despite the overhead involved in the extra grouping functionality, that user's performance will improve using the grouping functionality in the ViGOR system in comparison with an appropriate baseline system.*
- *Hypothesis H1.2: Users will explore more aspects of their task using ViGOR and that the workspace will help the users explore and see more options in large and unfamiliar datasets.*
- *Hypothesis H1.3: Users will be more satisfied with their search results and the search process using ViGOR.*

For the second evaluation which compared the basic ViGOR system with the ViGOR extension with recommendations, we had four hypotheses:

- *Hypothesis H2.1: The use of implicit information from previous users will help address the nosiness of implicit information on video retrieval systems.*
- *Hypothesis H2.2: Users conducting multi-faceted and ambiguous video retrieval tasks can benefit from recommendations based on implicit feedback.*
- *Hypothesis H2.3: The organisational features (i.e. the grouping functionality and the organisation of those groups in the workspace) available in ViGOR allow richer and multi-faceted recommendations (i.e. recommendations that incorporate multiple diverse videos that may be relevant to the search task).*
- *Hypothesis H2.4: The use of soft links in the implicit pool representation allows more useful and diverse recommendations.*

In the following sections we will describe both evaluations in full detail and will also outline the results obtained for both evaluations with respect to the hypotheses that have been outlined above.

5.2 Collection and Tasks

For the purposes of this evaluation we used the YouTube API to provide access to YouTube's video collection. In order to evaluate all systems, we made use of simulated work task situations [2]. For the first evaluation, which evaluated the ViGOR system, four simulated work task situations were created in order to provide broad, ambiguous, open ended tasks for the users for both evaluations. The second evaluation, which evaluated the ViGOR system with recommendations, also made use of four simulated tasks. Two of the tasks for the second evaluation were similar to tasks from the first evaluation, but for the first evaluation we asked the participants to search for videos related to specific aspects of the task, in addition to whatever other aspects they wished to investigate. The similar tasks for the second evaluation were broader as we removed the restriction on users which asked them to investigate specific aspects, although users were still asked to investigate at least three distinct aspects for each task. The other two tasks are what we refer to as "supersets" of two tasks from the first evaluation. By superset we mean that the relevant videos from the first task could be considered a subset of the relevant videos for the given task for the second evaluation, thus the new tasks for the second evaluation could potentially contain aspects related to tasks from the first evaluation. Following the completion of each of these tasks, the users were asked to write a short essay or similar about their results, in this way we could evaluate how many aspects of the task that the user investigated and how rich these aspects were, as well as getting an indication of the user's goals and aims for the retrieval task. An example of work task description is presented in Figure 5. The evaluated work tasks are outlined in terms of the

indicative search task below (where the task for the first evaluation is presented first and then the task for the second evaluation is presented):

Task 1 Politics

- Evaluation 1: A task of finding videos of political figures of 2008
- Evaluation 2: A task of finding videos containing leading world figures (Superset)

Task 2 Travel

- Evaluation 1: A task of finding video clips about Paris, Rome and other Europe locations
- Evaluation 2: A task of finding video clips about locations in Europe that you would like to visit (Similar)

Task 3 Culture

- Evaluation 1: A task of finding videos that illustrate Scottish culture, in particular Scottish dancing and food
- Evaluation 2: A task of finding videos that illustrate Scottish culture (Similar)

Task 4 World News

- Evaluation 1: A task of finding the major sport news of 2008
- Evaluation 2: A task of finding videos illustrating news stories from 2008 (Superset)

Task 1: Find Leading World Figures

Simulated Work Task Situation

Imagine that you are a student and as part of your class you must make a presentation about leading world figures from the early 21st century e.g. Saddam Hussein, George Bush, Tony Blair, Bono, Bill Gates etc. For the piece you must find various videos about each figure and write a short description of your presentation outlining the different aspects of the news and personality of that figure as shown by the clips that you have selected of that figure. Feel free to find shots of as many figures as you like and mark any videos that you feel are relevant.

*You may assume that editing software is available which will be used later on to edit the video clips that you find and select the best shots. The description will be written as part of a post task questionnaire.

Indicative Request

A task of finding videos containing leading world figures.

Figure 5. Simulated work task situation for Task 1 Politics (Evaluation 2).

For the second evaluation, which included recommendations, the same implicit pool was used for all tasks, created from the first evaluation. As we defined the similar/superset tasks on the second evaluation, users were not receiving recommendations from identical tasks. Also, as the tasks are extremely broad, many users conducted the tasks in very different ways, hence they would not repeat queries and/or interactions, and indeed may not have the same end goal in mind as the previous users from the first evaluation. The implicit pool was thus created from the implicit information of the 16 users from the first evaluation, with an average of 35 relevant documents retrieved and 4.5 groups created per task during the first evaluation. This implicit pool contained 5.5K nodes, 9.7K direct links (l=1) and ~7K soft links per each soft link distance level (up to l=10). We also included the usage information of the 8 users that interacted with the baseline interface which mimics YouTube (see Figure 5), as their information could be exploited by the global recommendation approach.

It should be noted that the second evaluation was conducted 5 months after the first evaluation; hence some changes in the YouTube index are to be expected and were accounted for in the following fashion. Firstly, the recommendation approach did not recommend items from the interaction information from the first evaluation that could no longer found by searching YouTube's index. Secondly, it was expected that new content relating to the tasks would be uploaded to YouTube between the two evaluations and as such this new content would not be included in the interactions from the first evaluation. This could potentially bias the search process of the users, as the recommendations would not include any content uploaded in the past 5 months. We hypothesised that this would be more noticeable in the similar-type tasks, as the superset-type tasks were related to finding information about to the previous year, and the available content would not have varied significantly between the two evaluations. In order to account for the variance between the YouTube index and our implicit pool, after every four users in the second evaluation we updated the implicit pool for the similar tasks. In this way, interaction with the new content would be added to the implicit pool for users conducting the similar tasks, but not for the superset tasks. In addition, this would allow us to

analyse and compare the effect of adding implicit information for the same exploratory tasks and using a static source of implicit information for related tasks.

5.3 Experimental Design

As has been mentioned previously two user centred evaluations were performed, during which three different systems were evaluated: a baseline system with mimicked YouTube functionality (YI, see Figure 6), ViGOR without recommendations (see Figure 1), and ViGOR with recommendations (see Figure 2). The first evaluation compared the baseline system with ViGOR in order to assess ViGOR's grouping paradigm. The second evaluation compared ViGOR with its extension with recommendations in order to assess the feasibility of our multi-faceted recommendation model. Two evaluations were conducted rather than one, as our first set of hypotheses address the utility of ViGOR as a video retrieval system and the second set address our recommendation approach. We wanted to make sure that ViGOR was an appropriate platform before using it to evaluate our retrieval approach. In addition the logs from the first evaluation provided a stable implicit pool for the recommendation algorithms in the second evaluation. A between subjects design was adopted for both evaluations. Each participant carried out four tasks in a Latin square design, using only one of the interfaces, which was randomly assigned. Upon arrival the participants were given an introductory sheet outlining the purposes of the experiment. If the participants were happy to proceed they then completed a consent form and an introductory questionnaire which gathered some background information on each participant. The participants were then given a demonstration of how to use the system. Following this training on the search system, the participants were allowed to complete a training task, this was to allow the participants to further familiarise themselves with the system and also see the types of tasks they had to complete for the evaluation. The same training task was used for all evaluations. Following this training the users began the evaluation; participants were allowed a maximum of 20 minutes to complete each of the four tasks. After each task the participants were asked to complete a post task questionnaire. In addition for each participant their interaction with the system was logged, as well as storing the videos they marked as relevant. Finally after they had finished all four tasks and post task questionnaires the participants were asked to complete an exit questionnaire. For both evaluations participants were paid a sum of £12 for their participation in the experiment, which in both cases took approximately 2 hours.

The first evaluation which compared the baseline YouTube system with ViGOR without recommendations, involved 16 participants, who were randomly divided into two groups of 8 and each group used one of the systems. The participants were mostly postgraduate students and researchers at out university. The participants consisted of 12 males and 4 females with an average age of 29 years (median: 27.5) and an advanced proficiency with English. The participants indicated that they regularly interacted with and searched for multimedia.

The second evaluation which compared ViGOR without recommendations with ViGOR with recommendations, involved 24 participants, which were randomly split into two groups of 12. The participants were mostly postgraduate students and researchers at our university. The participants consisted of 18 males and 6 females with an average age of 28.78 years (median: 28) and an advanced proficiency with English. Once again the participants indicated that they regularly interacted with and searched for multimedia. There was no overlap between the two groups of users.

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 hypotheses and the potential benefits of the systems is laid out in the following sections.

6. ViGOR Interaction Results

The first evaluation compared the performance of the ViGOR system (see Section 3.1) with a baseline system that mimicked YouTube's functionality, which will refer to as YouTube Interface (YI). ViGOR offers three expansion options for each group (see Figure 1 (c)): 1) related videos; 2) videos from the same user 3) and text expansion, which is the result of a new search using text extracted from the selected videos. All of the videos returned by these expansion options are retrieved using the YouTube API. The baseline interface, YI (see Figure 6) allowed users to search via text (A) and, when a video was

playing (B), users were presented with lists of related videos (C) and videos from the same user (D), in the same way that YouTube did at the time, this also mimicked the functionality available through the group expansions explained above. In addition, users of the YI were provided with a panel where they could drag and drop relevant videos (E). Similarly, users of ViGOR were instructed to organise relevant results in each group panel. Each participant carried out four tasks either using the YI or ViGOR. The results of the evaluation are presented below. In order to provide clarity we reiterate the hypotheses that are addressed by the findings at the beginning of each section, this presentation style is continued for each subsection of results.

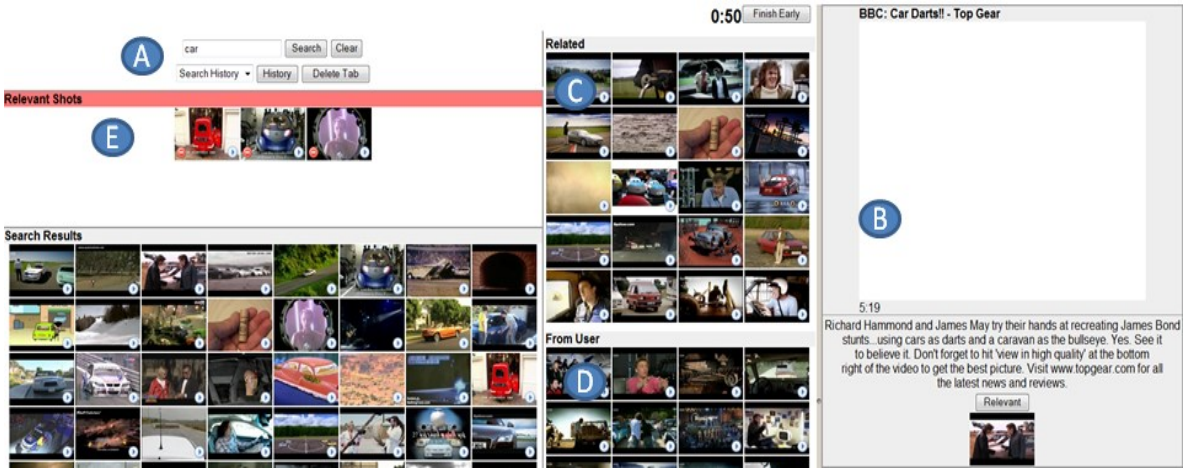

Figure 6. YouTube interface (YI).

6.1 System Performance

H1.1: Despite the overhead involved in the extra grouping functionality, that user's performance will improve using the grouping functionality in the ViGOR system in comparison with an appropriate baseline system.

H1.2: Users will explore more aspects of their task using ViGOR and that the workspace will help the users explore and see more options in large and unfamiliar datasets.

The results of analysis of interactions with the evaluated systems are shown in Table 1. The interactions were compared using a multivariate analysis of variance (MANOVA), the independent variables were system and topic, the dependent variables were use of tooltip, videos viewed, number of queries, number of videos marked as relevant, number of videos marked as irrelevant, number of aspects explored and time to complete the topic. A two-way MANOVA was used to account for Type 1 errors and to examine any potential interaction between topic and system. For the multivariate tests the interaction between topic and system was not found to be significant ($F(21,144.123)=0.915$, $p=0.573$; Wilks $\lambda=0.698$, partial $\eta^2=0.113$). Topic was not found to be a significant factor ($F(20,144.123)=1.201$, $p=0.259$; Wilks $\lambda=0.629$, partial $\eta^2=0.143$), but system was found to be a significant factor ($F(6,50)=3.506$, $p=0.004$; Wilks $\lambda=0.671$, partial $\eta^2=0.329$). Tests of between subject effects for system found that most differences were not statistical significant, except for the difference between the number of videos viewed in both systems ($F(1, 56)=17.064$, $p<0.001$, partial $\eta^2=0.234$), with more videos viewed using the YI (see Table 1 and Table 2), and the number of deleted videos (or marked as irrelevant), with ViGOR ($F(1, 56)=4.122$, $p=0.047$, partial $\eta^2=0.069$). Despite the lack of significant differences, there are some promising indicative trends worth highlighting. First, users of ViGOR marked 35.09 videos as being relevant (by assigning them to a group) per task in comparison with 23.16 videos for users of the YI ($F(1, 56)=2.830$, $p=0.098$, partial $\eta^2=0.048$). Second, this was also achieved in descriptively less time, with users of ViGOR completing their task in 18.6 minutes in comparison with 19.06 minutes for users of the YI ($F(1, 56)=0.064$, $p=801$, partial $\eta^2=0.001$). While it is clear that the addition of the grouping functionality does not negatively impact performance, further studies must be performed to conclude that the system is more effective, although an increase of the number of retrieved videos in less time is a potentially promising early indication of improved performance.

As discussed above, in absolute terms YI users viewed 19.9% more videos than ViGOR users (total of 25.1 videos for YI and 20.1 videos for ViGOR average per task), while other type of actions had no statistical difference (see Table 1). This means that users benefit from ViGOR by reducing the number of *high cost* interactions thus freeing the user to explore the task that they are trying to complete. We also note a trend in which, in absolute terms, there is an increase of *low cost* interactions from the use of the tooltip functionality and the selection of relevant/irrelevant results. The former is a lightweight functionality for the user to carry out, as it does not require to, e.g., play the video in the player or input text for a new query. The latter is an inherent action of finding more relevant results through the use of the evaluated system. While there is no statistical evidence of the increase on low cost interactions, we welcome these results, as some studies, e.g. [3], suggest this increment is a desirable outcome, as fostering a richer interaction between the user and the search system is a preferred goal when performing complex search tasks, as it allows for a better definition of the information need of the user.

Users of the ViGOR system created an average of 4.5 group panels, which shows that they went well beyond the three mandatory aspects and were comfortable using the interface to investigate a number of aspects. We evaluated the results and essays created by the users of the YI interface to determine the number of aspects that were investigated. Users of the YI interface created slightly less aspects than the ViGOR interface, 4.1 aspects on average. Although the results are not statistically significant for this sample size, it is worth noting the trend that users that were using ViGOR investigated more aspects of the tasks. Furthermore, investigating more aspects coupled with the higher number of relevant documents retrieved indicates a trend in which ViGOR users create richer, more complex and more detailed aspects than those of the YI.

	ViGOR		**YI**			
	#	**%**	**#**	**%**	**F**	**p**
Tooltip	79.4	61.2%	60.1	58.2%	3.444	0.069
**View*	5.1	3.9%	6.3	6.1%	17.064	0.000
Query	12.9	10.0%	14.1	13.7%	0.150	0.700
Relevant	31.7	24.4%	22.1	21.4%	2.830	0.098
Irrelevant (Deleted)	0.8	0.6%	0.6	0.6%	4.122	0.047
Aspects explored	4.5		4.1		0.006	0.940

Table 1. Total number of different interactions averaged per task and interface. Interactions are for unique videos e.g. if a user plays the same video twice we only record it once. Significant differences are marked with *.

Thus far the results indicate that users of ViGOR seem to explore the collection more and create more and richer aspects, with more related content found for each aspect. However, none of the differences in interaction or performance are significantly different, but the trend is that performance is improved when using the ViGOR interface over the YI. While the ViGOR users have more interaction with the retrieval system, they rely on less taxing actions such as tooltip, to the detriment of other more taxing actions such as playing a video for a specific duration or issuing a new search query. While the analysed trend is promising, further evidence is needed to support H1.1 and H1.2. In addition, the organisational facilities afforded provided by ViGOR bring about other benefits and interaction opportunities, without negatively impacting user performance. In an attempt to validate hypothesis H1.3, we analysed user feedback provided by the questionnaires that the participants completed at different stages of the evaluation.

6.2 User Feedback

H1.3: Users will be more satisfied with their search results and the search process using ViGOR.

In post search task questionnaires we solicited subjects' opinions on their assigned system and their reaction to the retrieved videos. The following 5-point Likert scales and semantic differentials were used. Some of these questions and terms are considered to be related and complementary but not identical. For example, by appropriate we mean a particular video is suitable for the task at hand, but may not necessarily be relevant, in this way appropriate is considered to be a slightly broader concept. "The videos that I have received through the searches were" "Relevant / Irrelevant" (Relevant), "Appropriate / Inappropriate" (Appropriate), "Complete / Incomplete" (Complete) and "Familiar / Strange" (Familiar). "I had an idea of

which kind of videos were relevant for the topic before starting the search" (Prior). "I found it easy to formulate queries on this topic" (Formulate). "During the search I have discovered more aspects of the topic than initially anticipated" (Discover). "The video(s) I chose in the end match what I had in mind before starting the search" (Match). "The tools provided allowed me to find videos that matched the topic" (Tools). "My idea of what videos and terms were relevant changed throughout the task" (Change). "I am satisfied with my search results" (Satisfy). The users were also asked about any issues that might affect performance on a scale from agree (1) to disagree (5), they were asked "What are the issues/problems that affected your performance" – "I didn't understand the task" (did not understand), "the video collection didn't contain the video(s) I wanted" (did not contain), "the system didn't return relevant videos" (no relevant), "I didn't have enough time to do an effective search" (no time), "I was often unsure of what action to take next" (no action) and "I found the system confusing" (confusing). In a post experiment questionnaire the two interfaces were compared using the following 5 point semantic differentials regarding overall reaction to the system with a positive response being higher on the scale – "wonderful/terrible", "satisfying/frustrating", "stimulating/dull", "easy/difficult", "flexible/rigid", "efficient/inefficient", "novel/standard" and "effective/ineffective". Table 2 presents the average responses for each of these scales using the labels after each of the Likert scales in the list above.

Differential	***YI***	***ViGOR***	***U***	***z***	***p***
Relevant	4.031	**4.094**	485	-0.401	0.689
Appropriate	3.875	**4.094**	434.5	-1.118	0.263
Complete	3.129	**3.375**	445.5	-0.953	0.340
Familiar	**3.906**	3.500	401	-1.573	0.116
Prior	**4.000**	3.875	478	-0.494	0.621
Formulate	3.812	**4.156**	504	-0.117	0.907
Discover	2.875	**3.312**	416.5	-1.354	0.176
Match	**3.718**	3.656	405.5	-1.467	0.142
Tools	3.844	**4.187**	482.5	-0.446	0.656
Change	2.687	**2.875**	424	-1.264	0.206
Satisfy	3.656	**3.750**	463	-0.682	0.495
Did not understand	4.780	**4.910**	464.5	-1.178	0.239
Did not contain	3.812	**4.500**	318.5	-2.848	0.004
**No relevant*	3.719	**4.500**	279	-3.348	0.001
No time	3.562	**4.312**	353	-2.310	0.021
No action	4.062	**4.437**	379	-1.935	0.053
Confusing	4.344	**4.656**	430.5	-1.294	0.196
Wonderful	**3.630**	3.500	28	-0.488	0.721
Satisfying	**3.880**	**3.880**	30.5	-0.167	0.867
Stimulating	**3.500**	3.130	22	-1.195	0.232
Easy	**4.250**	**4.250**	32	0	1
Flexible	2.750	**3.880**	13.5	-2.019	0.044
Efficient	3.630	**3.750**	28.5	-0.400	0.689
Novel	2.130	**3.380**	10.5	-2.380	0.017
Effective	3.870	**3.880**	30.5	-0.177	0.860

Table 2. Perceptions of Retrieved Videos (Higher = Better). The most positive response is in bold and significant differences are marked with *. Degrees of freedom for each are 32.

The questionnaire responses were compared using a Mann Whitney U test with system as the independent variable with a Bonferroni adjusted alpha of p=0.002. With respect to their perceived performance, users perceived that when using YI that the videos returned were not relevant (U(32)=279, Z=-3.348, p=0.001). Overall from the results in Table 3 it appears that participants have a better perception while interacting with ViGOR, as the trend is that the give more positive responses for ViGOR on comparison with YI, however as with H1.1 and H1.2 most of these differences are not significant. These findings

indicate that users will be more satisfied with their search results and the search process using ViGOR, while the users are more positive about ViGOR than the YI they are not significantly so. Thus hypothesis H1.3 is not fully supported, however the addition of the grouping functionality does not negatively impact user perception and indeed as noted the trend is for a slight increase in user satisfaction.

6.3 Summary

In this section we have evaluated the impact of the ViGOR interface on video search. ViGOR was compared to a baseline interface resembling a familiar video search paradigm as provided by YouTube. The main novelty of ViGOR is the addition of a workspace that allows users to group results, which we believe facilitates the exploration of specific aspects of a broader search task. The evaluation findings, although not definitive, provide some support for our hypotheses, suggesting that, although the interaction with the workspace adds an additional overhead to users' search process, the creation of aspects was perceived by many participants as a more flexible way of approaching exploratory search tasks. What could be viewed as the main drawback of this new search paradigm is the increase on number of interactions required by users. However, it should be noted that a trend in ViGOR was that users spent more time interacting with the workspace and results (e.g., through tooltip actions) than issuing keyword queries or viewing videos. This suggests that the users were comfortable with the new search paradigm, and, as our system performance study indicated, there was no negative effect on the search performance of users when using ViGOR. There is even an early indication that users could potentially perform better when using the organisational features made available by ViGOR, but further studies have to be conducted to validate this claim. Nevertheless, as we show in the next section, ViGOR enables richer user interactions when performing exploratory search tasks, which in turn can be used to produce multi-faceted recommendations.

7. ViGOR Recommendation Results

The second evaluation compared the ViGOR system without recommendations used in the first evaluation with a ViGOR system extended with recommendations. As explained in Section 3.2, ViGOR with recommendations shows a global panel of video recommendations, above the search result panel (see Figure 2(E)), which proactively changed with every interaction of the user with the system, using the global recommendation approach introduced in Section 4.4. The extended version of ViGOR with recommendations also has an added new local expansion option; which uses the local recommendation approach introduced in Section 4.5.

7.1 Task Performance

H2.1: The use of implicit information from previous users will help address the nosiness of implicit information on video retrieval systems.
H2.2: Users conducting multi-faceted and ambiguous video retrieval tasks can benefit from recommendations based on implicit feedback.

In order to investigate hypothesis 2.1, we performed a direct comparison using a MANOVA, the independent variables were system and topic, the dependent variables were use of tooltip, videos viewed, number of queries, number of videos marked as relevant, number of videos marked as irrelevant, number of aspects created, number of aspects deleted, user expansion, related expansion, text expansion and time to complete the topic. For the multivariate tests the interaction between topic and system was not found to be significant ($F(36,228.233)=0.825$, $p=0.752$; Wilks $\lambda=0.697$, partial $\eta^2=0.113$). Topic was not found to be a significant factor ($F(36,228.233)=1.043$, $p=0.410$; Wilks $\lambda=0.638$, partial $\eta^2=0.139$). System was found to be significant ($F(12,77)=8.603$, $p<0.001$; Wilks $\lambda=0.427$, partial $\eta^2=0.573$). The analysis of user actions executed is shown in Table 3, the global recommendation is not shown as it was automatically updated. Tests of between subject effects for system showed that system affected the number of videos marked as relevant (i.e. videos that were in groups at the end), it was found that on average users of recommendation system marked 28.52 videos as being relevant per task in comparison with 20.191 videos for users of the system without recommendations ($F(1,88)=5.786$, $p=0.017$, partial $\eta^2=0.063$). In addition to this, users of the recommendation system created more groups or aspects of the task on average, 5.604, as opposed to

4.702 for the system without recommendations, the difference between systems was significant ($F(1,88)=4.616$, $p=0.034$, partial $\eta^2=0.050$). A higher number of created groups indicate that users explored more aspects of each particular task. Overall, these results show that users are retrieving more videos and expressing more aspects of their task using ViGOR with recommendations.

The values highlighted in Table 3 show that the users of ViGOR with recommendations have more user interactions with the system overall in comparison with users of the ViGOR baseline system. Much of this difference is due to the increased use of the tooltip functionality of the recommendation system users; this is a lightweight functionality which is of low cost for the user to carry out. In terms of more heavyweight user actions such as querying the system or viewing a video, there are small differences between the two systems. Users of the recommendation system seem to submit slightly more queries and view more videos. However, the differences above in these values are not significant. One major noticeable difference in the user interactions is the way that the users use the expansions. In both systems the expansion by more videos from the same user and expansion by text are not used very often, however the difference in expansions by the same user is significantly different between the two systems ($F(1,88)=11.775$, $p=0.001$, partial $\eta^2=0.118$) with users using the feature less in the system with recommendations. In contrast the query by related video from YouTube is the most frequently used expansion. In the recommendation system we see that the three YouTube related expansions are used less frequently than in the baseline ViGOR system. This is to be expected as this system has one more local expansion option. However, it can be seen that the new recommendation expansion is used almost as frequently as the YouTube related expansion. This is an encouraging result; all of our users had previous experience using YouTube and were familiar with related videos etc., but not the recommendations, users appear to find the recommendations quite useful and exploit this resource. As was stated in the experimental design section (see Section 5) users also filled out questionnaires after each task. The analysis of the user responses did not indicate a preference towards either system.

Action	ViGOR	ViGOR + Recommendations	F	p
Tooltip	79.723 (43.026)	97.208 (52.245)	3.204	0.077
View	33.659 (24.143)	39.812 (51.878)	0.490	0.486
Query	14.766 (8.215)	15.667 (10.035)	0.320	0.573
**Add to group*	20.191 (8.363)	28.520 (18.007)	5.876	0.017
Delete from group	1.787 (3.444)	1.416 (2.019)	0.403	0.527
**Create group*	4.702 (1.966)	5.604 (2.377)	4.616	0.034
Delete group	0.234 (0.520)	0.333 (0.595)	0.879	0.351
Expand Text	1.957 (2.245)	1.542 (2.083)	1.371	0.245
**Expand User*	1.553 (1.755)	0.625 (0.890)	11.775	0.001
Expand Related	5.809 (4.292)	4.458 (4.708)	2.567	0.113
Local Recom.	n/a	3.354 (3.479)	n/a	n/a
Time (minutes)	17.73 (2.92)	18.46 (2.02)	1.623	0.206

Table 3. Average number of interactions per task for each interface, standard deviation in brackets. Significant differences are marked with *.

Thus far we have seen that the user performance in terms of videos retrieved improves significantly with the use of the recommendations, in addition users also investigate slightly more aspects of the task, by creating more group panels. These findings provide partial validation for hypothesis H2.1 and H2.2, as the systems was able to exploit past noisy implicit information to benefit the users in their explorative and multi-faceted tasks in terms of found relevant videos and explored aspects. A trend was found, however, in which user interactions increase while using the recommendation system. Even so, most of this increase is due to an increase in the use of the lightweight tooltip function, this may just be as a result of the extra results and options that are presented to the users of the recommendation system. In the next section we will investigate this user behaviour in more detail.

7.2 User Interactions

H2.3: The organisational features (i.e. the grouping functionality and the organisation of those groups in the workspace) available in ViGOR allow richer and multi-faceted recommendations (i.e. recommendations that incorporate multiple diverse videos that may be relevant to the search task).

In order to investigate H2.3 and also to investigate previous findings further, the user behaviour while using the recommendations was analysed in more detail. Table 4 shows the average number of videos selected from each expansion or recommendation. The only prior knowledge users had about all expansion options was based on their previous interactions with YouTube. It can be seen clearly that the text expansions and the videos from the same user expansions return the videos that the users selected least often; this indicates that the users did not find these types of recommendation useful. In terms of the most used technique, the YouTube related expansion is the most used for the two superset tasks (Tasks 1 and 4). The global recommendations are the most used technique for the two similar tasks (Tasks 2 and 3). Similar and Superset tasks are designed to investigate different situations of available usage information: superset tasks make use of a static implicit pool based on usage information from previous more specific tasks, whereas the similar tasks made used of previous information from similar tasks and also were updated with the information obtained from previous users during the evaluation. Due to the nature of the tasks and the experimental design (see Section 5.2) it is more likely that users will search for similar aspects of the task in the similar tasks than in the superset tasks, thus this tasks are more likely to receive more relevant recommendations. However, for the superset tasks, the local and global recommendations are still useful for the user, as they are used more than the baseline expansion approaches such as the user expansion and the text expansion.

Table 5 shows the percentage of selected videos that came from each local expansion or recommendation functionality, over the selected videos from all these approaches, without considering selected videos from the textual search panel. This gives a relative value of importance for each of the expansion and recommendation approaches, the greater this value, the more the users used this recommendation option when executing their search tasks. We can see that the global recommendation has more importance on average that the YouTube related local expansion. In fact, the importance of the YouTube related function drops on average from an 85% to a 38% from ViGOR to ViGOR with recommendations. This seems to indicate that users rely less on this feature, which is dominant on the baseline system, and prefer to utilise the newer recommendation approaches. The global recommendation approach seems to be much more important globally than the local recommendation approach, which has similar values as the textual based expansion. Whereas Table 4 showed a greater difference in absolute terms on the selection values of the recommendation on the similar and superset tasks, Table 5 shows that the importance of the recommendation approaches are only slightly higher on the similar tasks. Moreover, regarding the overall performance of these tasks, it is important to note that the increase in performance in comparison with the baseline system is consistent across all tasks. For instance, there was an average increase of 39.44% relevant documents on the similar tasks and an average increase of 37.53% relevant documents on the superset tasks.

	YouTube API Expansion			**Recommendation**		
Task (Type)	**Text**	**YouTube Related**	**YouTube User**	**Local Recom.**	**Global Recom.**	**Total**
1 (Superset)	0.25 (2.92)	6.58 (6.50)	0.00 (2.33)	0.08	1.50	8.42 (11.75)
2 (Similar)	0.42 (1.83)	3.67 (6.33)	0.42 (1.08)	1.25	4.42	10.17 (9.25)
3 (Similar)	0.75 (1.77)	3.83 (6.23)	0.00 (1.31)	2.83	5.42	12.83 (9.31)
4 (Superset)	0.50 (1.83)	2.58 (4.58)	0.17 (1.75)	0.50	0.92	4.67 (8.17)
Avg	0.48 (2.08)	4.17 (5.92)	0.15 (1.61)	1.17	3.06	9.02 (9.61)

Table 4. Number of documents selected on average per user and search task for each expansion or recommendation approach: text related expansion (Text); YouTube expansion by related videos (YouTube Related); YouTube expansion by videos uploaded by the same user (YouTube User); local recommendation approach (Local Recom., Section 4.5); global recommendation of videos (Global Recom., Section 4.4); total average selection of videos for all recommendation and expansion options (Total). Values in brackets indicate values from the baseline system, where appropriate.

	YouTube API Expansion			Recommendation	
Task (Type)	**Text**	**YouTube Related**	**YouTube User**	**Local Recom.**	**Global Recom.**
1 (Superset)	10.19% (14.92%)	45.57% (78.75%)	0.00% (6.33%)	2.78%	41.47%
2 (Similar)	9.74% (6.89%)	36.19% (91.24%)	1.89% (1.88%)	5.98%	46.20%
3 (Similar)	10.78% (3.06%)	27.03% (90.61%)	0.00% (6.34%)	12.07%	50.13%
4 (Superset)	7.30% (10.87%)	46.63% (76.77%)	2.86% (12.35%)	10.60%	32.62%
Avg	9.52% (8.52%)	38.34% (85.05%)	1.20% (6.43%)	8.04%	42.90%

Table 5. Percentage of incoming local expansion and recommendation source for selected videos as relevant for each expansion or recommendation approach: text related expansion (Text); YouTube expansion by related videos (YouTube Related); YouTube expansion by videos uploaded by the same user (YouTube User); local recommendation approach (Local Recom., Section 4.5); global recommendation of videos (Global Recom., Section 4.4); total average selection of videos for all recommendation and expansion options (Total). Values in brackets indicate values from the baseline system, where appropriate.

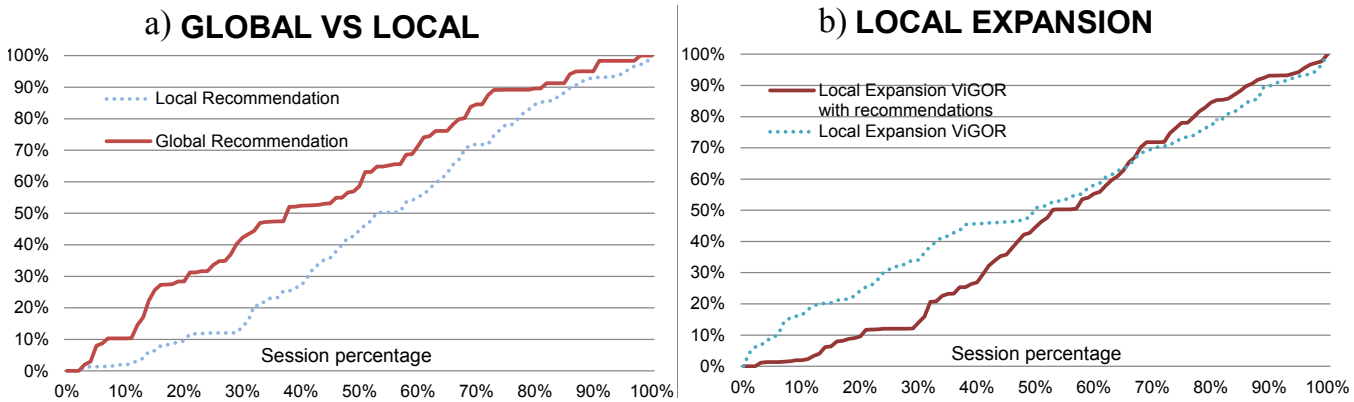

Figure 7. a) Cumulative distribution of selection of recommendations over session percentage completion. b) Cumulative distribution of local expansion execution over session percentage completion.

In an attempt to gain a further insight into the differences between the user interactions with the two types of recommendation approaches, we plotted a cumulative distribution of the execution of each type of recommendation against session completion (see Figure 7a). We do not show any of the other three expansions actions in this figure, as they follow a similar distribution to the local recommendation. A pair wise t-test revealed that the differences between the recommendation distributions were statistically significant. Figure 7a shows that users select examples from the global recommendations early in the task. It is not until later in the task that the users appear to select examples from the local expansions and add them to groups. Figure 7b illustrates this change on user behaviour between ViGOR and ViGOR with recommendations. The figure shows the cumulative distribution of all local expansion functionalities (including the local recommendation on ViGOR with recommendations). We can see that there is a difference on how the local expansions are executed on each system, as users of ViGOR with recommendations rely less on the local expansion on the first stages of the search.

While the results highlighted by Table 5 suggest that the impact of our recommendation approach is similar across all tasks, the values shown in Figure 7 illustrate a difference in user behaviour. It appears that at the beginning users are more interested in the overall global task, but as the task progresses users become more interested in the details of each aspect, relying in local expansion options. This supports our hypothesis H2.3, as this analysis shows how different types of recommendation have assisted users in different stages of their task.

7.3 Use of Soft Links on Global Recommendation

H2.4: The use of soft links in the implicit pool representation allows more useful and diverse recommendations.

To address hypothesis H2.4, a simulated analysis was performed. As described in Section 4.4, one of the recommendation approaches presented in this paper, the global recommendation, is an extension of a recommendation approach developed previously [21]. We can thus investigate if our extension results in a more effective recommendation approach, thus giving more insight over our hypothesis relating to user performance. The previous recommendation approach was extended by its adaptation to the grouping paradigm available in ViGOR and the use of soft links. The former extension was necessary in order to take advantages of the new interaction model offered by ViGOR, the main goal of the latter extension, soft links, is to be able to discern between the multiple facets that occur during a typical search task executed in ViGOR.

In order to test our hypothesis, we utilised the data obtained from the second user evaluation, which was described above. Using the interaction information obtained from the experiment, we followed the simulation framework presented by Vallet et al. [34]. One of the advantages of using this evaluation framework is that we can compare our proposed recommendation approach with the previous approach without the need of performing an additional user evaluation, which usually requires a significant amount of resources. Furthermore, the simulation approach will allow us to investigate the effect of soft links and the soft link level parameter L. The obtained results, however, cannot be compared with a user evaluation, but it allows us to tackle our research question in a resource-efficient way.

The simulation framework exploits the interaction information obtained in an interactive study in order to simulate users performing a search task. This framework uses the interaction information in two ways: 1) to represent a pool of collected user interaction information from previous users, as described in Section 4, which serves as training data for the evaluated recommendation approaches; 2) to provide statistical data in order to simulate a user using the search system, such as the probability that a relevant video is clicked or tooltipped, viewed for a specific time, added to a group, etc. The simulation framework was then used to obtain a number of candidate recommendations provided by both the baseline recommendation and our new global recommendation approach. Each recommended result was manually assessed in order to judge 1) if the result was relevant to the search task and 2) to which topic aspect this result may be related, regarding the current search topic (e.g. for Task 1, Politics, each world figure was assigned to a different aspect).

In order to evaluate the accuracy of the recommendation approaches, we used relevance judgments to compute the accuracy of our recommendations (P@N, average percentage of relevant documents in the top N results). In order to evaluate how diverse were the recommended results, we used aspect judgments in order to compute how many distinct aspects related to the search query and relevant results were recommended (Aspects@N, average number of distinct aspects present in the top N results). This relevance and aspect judgment was performed by 5 members of our research group, using a dedicated evaluation interface. The total assessment took approximately 1 hour of work for each assessor, resulting in a final judgement pool of 1535 results.

L	P@5	P@10	P@15	Aspects@5	Aspects@10	Aspects@15
Baseline	43.9%	54.4%	61.7%	2.026	2.503	2.933
3	54.4%*	65.7%*	69.5%*	1.954	2.503	3.217*
5	59.1%*	65.6%*	66.0%	2.313*	2.634	3.274*
8	60.6%*	67.0%*	68.1%*	2.267	2.726*	3.457*
10	67.5%*	66.8%*	69.2%*	2.518*	2.845*	3.623*

Table 6. Evolution of number of aspects for varying values of soft link level (L). Starred values indicate a statistical significant difference in comparison to the baseline method, which does not use soft links (Wilcoxon, $p < 0.05$).

Table 6 shows the results of the evaluation. The overall performance values, given by the P@10 metric, are around 55-67% for any value of L, which is a good accuracy value for a recommendation algorithm. The baseline approach, which is the approach presented by Hopfgartner et al. [21], has lower precision values than the soft link approach with increasing values of L, with statistically differences at various levels of soft links. In terms of P@10, the use of soft links result on an improvement of at least 20.7% over the baseline. This is an encouraging result as there is normally a trade-off between diversity and accuracy. Other interesting results are those related to the diversity of the results, measured by Aspects@10 and Aspects@15. It can also be seen in Table 6 that as the level of soft link L increases, the number of aspects presented in the

recommendation increases as well. For instance, our soft link approach with a value of L=10 presented on average 13.7% more aspects on the top 10 recommended documents (Aspects@10) than the baseline approach. We tested these differences on aspect diversity and they were found to be statistically significant (starred values in Table 6, Wilcoxon test, $p < 0.05$).

The results obtained in this section validate hypothesis H2.4, that the use of soft links in the implicit pool representation allows better and more diverse recommendations, as the results indicate that when increasing the maximum level of soft links considered. That is that there is a positive effect on the quality and the diversity of the recommendations, in comparison with the previous baseline approach [21]. Hence, our soft link approach achieves a higher diversity of results than the baseline approach, without a negative impact on the relevance of the recommendations.

7.4 Summary

In this section we have investigated the use of recommendation approaches within (and in some ways as an extension to) the workspace paradigm presented in ViGOR. As this new search paradigm allows for a richer interaction for the user during exploratory search, our hypotheses were focused on the impact of the new recommendation approaches over how users were interacting with the workspace. One of the main findings was that users seemed to rely on different recommendation approaches at different stages of their search. The way in which the recommendation approaches were used by users suggests that the search methodology adopted by the users was split in two stages: a first stage where users investigated different aspects related to the current search task (this is when they used more global recommendations), and a second stage, where users focus more on each specific aspect (this is when they used more local recommendations). Our global and local recommendations seemed to provide support for the first and second stage, respectively. This interesting finding also has a number of implications for exploratory video search and indeed providing results and recommendations for search in a number of different paradigms. One implication would be to tailor the recommendation approaches to better favour different stages of the search process. As the first stage of this search process seems to be related to discovery of new aspects, we also investigated the use of soft links as an extension to the global recommendation approach, our results from a simulated study indicated that soft links resulted in more diverse results as well as more accurate results, thus allowing users to more easily find other relevant aspects related to their search.

8. Discussion

In this section we provide a summary of our results, as well as discussion of some of the wider implications of our findings.

8.1 Grouping functionalities applied to video retrieval

One of our goals in this paper was to investigate three hypotheses relating to the use of ViGOR: H1.1) that user performance would improve through the use of ViGOR; H1.2) that ViGOR can aid user exploration of the task at hand; and H1.3) that the use of ViGOR can also increase user satisfaction with their search and their search results. To that end we conducted a user evaluation involving 16 participants, on a set of exploratory video search tasks that incorporate different user goals. While very few significant differences were found in terms of interaction between ViGOR and the baseline there are a number of interesting points that can be made about the results of these evaluations. It was found that the use of the grouping functionality resulted in users being as effective as with the standard interface. A trend was observed in which users retrieved more search results in comparison with a baseline system. While this difference was not found to be significant it is still a promising, indication of the benefits of the grouping approach. Our observation on the number of retrieved videos was also coupled with an increase in user interactions. However, most of the interaction increase can be attributed to non-expensive lightweight functionalities, while more expensive heavyweight functionality decreases, in comparison with the baseline interface, users of the grouping interface viewed 18% less videos and carried out 5% less queries. Our analysis also indicated that these results were also retrieved by these users in less time than users of the baseline system, although

differences were not statistically significant. Overall while the first two hypotheses were not fully validated the availability of the grouping functionality did not harm user performance when searching digital video libraries, while at the same time requiring them to view less videos.

There were also a number of interesting findings in terms of user perceptions with respect to their search process and the search system when using ViGOR. The trend was that users had a preference for ViGOR. Whereas in favour of the baseline the trend was that users found the baseline easier to use and easier to learn to use, they still had a preference for the grouping interface and found it to be better overall. The only significant difference in feedback for this comparison was that users had the perception that the baseline retuned less relevant videos. Although users could not directly compare the interfaces, users did highlight that their perception of their search was better when using the ViGOR system. These results show a trend that the users had a preference for the interfaces that provide the grouping functionality.

Overall it can be seen that the addition of grouping functionality for video search tasks could lead to a number of favourable outcomes, while there were very few significant differences between the performances while using both interfaces, the performance of participants did not decrease with the addition of grouping in the more complicated ViGOR interface. There are also a number of additional benefits that occur as a result of using a grouping search metaphor. The interactive grouping is a supple means of communicating a multitude of information needs e.g. short-term vs. long-term, specific vs. multi-faceted. The semantic gap is narrowed by the abstraction to high-level semantic groupings, reflecting an individual's task-specific mental model of the data. In addition as this grouping functionality operates at an interface level, this grouping paradigm can be applied to datasets and systems on large scales, as it can sit on top of any existing search system. Finally, the user leaves a trail of their interactions, which can not only be exploited by the system for adaption but by which can be traced by other users for collaboration. This was exploited in our second evaluation, which is discussed below.

8.2 Multi-faceted recommendations applied to video retrieval

The concept of ViGOR was designed to allow users to conceptualise their search task by creating groups of videos to solve a video search task. To build upon this a new recommendation approach based on a concept of soft links was integrated with ViGOR (exploiting the trails highlighted in the conclusion of the section above), this recommendation approach was used to provide recommendations that are based on implicit feedback. The unique organisational features available in ViGOR allow for richer and multi-faceted recommendations to help users at different points in their search process. In order to evaluate our recommendation approach, we defined a number of hypothesis: H2.1) The use of implicit information from previous users will help address the nosiness of implicit information on video retrieval systems; H2.2) Multi-faceted and ambiguous video retrieval tasks can benefit from recommendations based on implicit feedback; and H2.3) The organisational features (i.e. the grouping functionality and the organisation of those groups in the workspace) available in ViGOR allow richer and multi-faceted recommendations.

The results of our evaluation showed that the users using the recommendation system retrieved almost 40% more videos in comparison to the baseline system. This increase came about with a minor increase in the effort that the user had to expend in order to find these videos. The majority of the increase in effort can be attributed to lightweight search features that are not that costly to the user in terms of time and effort. In fact the users of the recommendation viewed slightly less videos and executed more searches, illustrating a further benefit as users were not having to make as many judgements about individual videos freeing them to explore further results. All of these results illustrate that users engaged in multi-faceted and ambiguous video retrieval tasks can benefit from recommendations based on implicit feedback. Also, it appears that the recommendations are overcoming the problems associated with inherent noise in implicit feedback for video search tasks. Furthermore, in order to tests the benefits of our soft link representation, we analysed the interaction logs from the user study and applied a simulation based methodology. The results of this analysis highlight that the representation and exploitation of soft links result on more diverse, yet relevant, recommendations, which can be considered a desirable quality for a recommendation approach that is applied to an exploratory search task over a large document collection. These findings thus provide some support for hypotheses H2.1 and H2.2.

In relation to hypothesis H2.3, the user interactions with the system were investigated in more detail. An initial examination revealed that the global recommendations were used most often by the users in comparison with localised specific recommendations. Further investigation of the user interactions showed that, as well as global recommendations being used more often than local recommendations, these types of recommendation were in fact used at different stages of the tasks by the users. The global recommendations were used more often at the beginning of the evaluation; whereas the local recommendations and expansion functionality were used at later stages in the search process. This finding could have implications for aiding the user search process in a variety of situations that involves multi-faceted and broad search tasks. It appears that at the beginning of the search process users are more interested in the overall task and how that task can be decomposed, thus they benefit from the diverse and broad global recommendations. At a later stage of their tasks users have already began exploring certain aspects and appear to be more interested in those specific aspects of their task.

An additional hypothesis, H2.4, investigated the recommendation approach that was adopted in this work. In order to test the hypothesis, we performed a simulation based study over the interaction information obtained during the user study. This analysis proved that the recommendation model benefits from use of soft links, by providing more diverse and higher quality recommendations.

9. Conclusions and Future Directions

In this paper we have introduced the ViGOR system, a video search and retrieval system that allow users to create groups of video search results to help conceptualise and organise their results for complex video search tasks. It was hoped that grouping search results on the workspace would motivate the user to organise results for their search/work task. This should enable the users to break up their overall search task into a small set of individual search tasks. Although the concept of grouping has been investigated in a number of retrieval scenarios [24] [33], its application and usefulness for searching video collections and archives is as yet not fully understood. As has been discussed previously (see Section 2.2), video provides a number of unique problems for search and retrieval that are not present in other search scenarios. Taking advantage of the new grouping and interaction functionalities that ViGOR offers, we also present a new multi-faceted recommendation approach, which is integrated into an extension of the ViGOR system. We believe that the combination of the grouping and recommendation functionality proposed in this paper can be applied successfully to a wide range of video collections and search frontends. It is worth noting that, although our experimental study is performed over a specific search collection, both the grouping and recommendation functionality does not require any special requirements from an online collection. This is exemplified by the translation of our grouping functionality from previous work on an offline collection such as TRECVID, which used our own indexing techniques [15], to a vast online collection such as YouTube, which uses the public search API. This was done without need of changing the interface or search frontend, and only adapting the traditional search features to each collection. Analogously, the recommendation techniques do not require any form of content representation, as they are solely based on gathering the implicit data from users. Content is solely identified by a unique ID, which is used to populate a pool of implicit information that feeds the different recommendation algorithms that produce as output a list of recommended content IDs associated to a score.

The benefits of adapting a search interface to the techniques studied in this work are varied. First, our initial study indicates that the grouping functionality allows richer interaction with the search system, without impacting the performance for exploratory search over vast multimedia collections. Although most of the obtained results are not significant, we identified a promising trend indicating that the impact of the grouping functionality alone could be beneficial. Nevertheless, we believe that the grouping paradigm allows users to better define their information need and to better structure the obtained results, thus they can focus on their high level needs rather than lower level tasks e.g. query formulation. It is worth noting that the evaluated system relied on user generated annotations, although in previous early work the grouping functionality also proved to be beneficial with low level features [15]. Second, our study on recommendation techniques based on implicit feedback indicated that the grouping paradigm can be complemented effectively with recommendation approaches. Our validated hypothesis indicate that, when performing explorative tasks, we may use recommendation approaches to both 1)

help on the first steps of the search, by broadening the initial user's concept of the search task at hand; and 2) help on expanding the exploration of a specific aspect of the search task, by offering recommendations that are applied to a single aspectual group.

Our findings, however, are focused on explorative and multifaceted search tasks. Although we believe that some of the approaches presented here could be applicable to other search paradigms, such as ad-hoc search (e.g. single result search) or browsing, further analysis has to be conducted in order to test the suitability of our approaches in such search contexts. We plan to analyse this suitability, and to study possible adaptations of the proposed techniques to such search paradigms. Our post-evaluation analysis has also highlighted that explorative search may have two different search phases, namely the *explorative* and the *focus* phase. In the initial *explorative* phase, users tend to make an initial broad search exploration of the different aspects involved in the task at hand. In the *focus* phase, users tend to explore in depth each of the aspects found in the *explorative* phase by looking for more related results. Although our initial experiments hinted that this was the usual way of performing explorative searches with the grouping functionality, we have a more definitive indication of this search tactic based on the experiments conducted for this work. Hence, we feel that the proposed recommendation approaches could be further tailored towards this search mechanism, by, e.g., supporting users with this behaviour and providing more help on this process.

In conclusion, the results of this evaluation have highlighted the promise of multi-faceted recommendations for video search tasks. Our recommendation approach based on implicit feedback coupled with an innovative video search interface has improved user performance and highlighted the promise of multi-faceted recommendations based on collaborative implicit feedback, for alleviating many problems associated with online video search, and indeed could be applied to numerous other video search paradigms.

10. REFERENCES

[1] Bauer, T., & Leake, D. B. (2001). Real time user context modeling for information retrieval agents. In *Proceedings of the tenth international conference on Information and knowledge management*. *CIKM '01*. (pp. 568-570).

[2] Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research* 8(3), paper number 152 .

[3] Bystrom, K., & Jarvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing and Management* 31(2), 191-213.

[4] Campbell, I. (2000). Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *Information Retrieval* 2(1), 89-114.

[5] Christel, M., & Conescu, R. (2006). Mining novice user activity with trecvid interactive retrieval tasks. In *TRECVID Interactive Retrieval Track.* (pp. 21-30).

[6] Christel, M. G. (2007). Establishing the utility of non-text search for news video retrieval with real world users. In *Proceedings of the 15th international conference on Multimedia. MULTIMEDIA '07.* (pp. 707-716).

[7] Craswell, N., & Szummer, M. (2007). Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. *SIGIR '07. (*pp. 239-246).

[8] Dou, Z. , Song, R. , & Wen , J. (2007). A large-scale evaluation and analysis of personalized search strategies. in *Proceedings of the 16th international World Wide Web conference. WWW2007.* (pp. 572-581).

[9] Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White R. (2005). Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2), 147-168.

[10] Fass, A. M., Bier, E. A., & Adar E., (2000). Picturepiper: using a re-configurable pipeline to find images on the web. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. *UIST '00.* (pp. 51-62).

[11] Fogarty, J., Tan, D., Kapoor, A., & Winder, S. (2008). Cueflik: interactive concept learning in image search. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. *CHI '08*. (pp. 29-38).

[12] Girgensohn, A., Shipman, F., Wilcox, L., Turner, T., & Cooper, M. (2009). MediaGLOW: organizing photos in a graph-based workspace. In *Proceedings of the 13th international Conference on intelligent User interfaces*. IUI '09. pp. 419-424).

[13] Guy, M., & Tonkin, E., Folksonomies. (2006). Tidying up tags?. *D-Lib Magazine*, 12(1).

[14] Halvey, M. J., & Keane, M. T. (2007). Analysis of online video search and sharing. In *Proceedings of the 18th conference on Hypertext and hypermedia*. *HT '07*. (pp. 217-226).

[15] Halvey, M., Vallet, D., Hannah, D., & Jose, J. M. (2009). ViGOR: a grouping oriented interface for search and retrieval in video libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. *JCDL*. (pp. 87-96).

[16] Hancock-Beaulieu, M. H., & Walker, S. (1992). An evaluation of automatic query expansion in an online library catalogue. *J. Doc.*, 48(4), 406-421.

[17] Hauptmann, A. G., & Christel, M. G. (2004). Successful approaches in the trec video retrieval evaluations. In *Proceedings of the 12th annual ACM international conference on Multimedia*. *MULTIMEDIA '04*. (pp. 668-675).

[18] Hauptmann, A. G., Lin, W. H., Yan, R., Yang, J., & Chen, M. Y. (2006). Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of the 14th annual ACM international conference on Multimedia*. *MULTIMEDIA '06*. (pp. 385-394).

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

[20] Hopfgartner, F., Urban, J., Villa, R., & Jose, J. (2007). Simulated testing of an adaptive multimedia information retrieval system. In *International Workshop on Content-Based Multimedia Indexing*. *CBMI 2007*. (pp. 328-335).

[21] Hopfgartner, F., Vallet, D., Vallet, & Jose, J.M. (2008). Search trails using user feedback to improve video search. In *Proceeding of the 16th ACM international conference on Multimedia*. *MULTIMEDIA '08*. (pp. 339-348).

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

[23] Mei, T., Yang, B., Hua, X. S., Yang, L., & Yang, S. Q. (2007). Videoreach: an online video recommendation system. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. *SIGIR '07*. (pp. 767-768).

[24] Nakazato, M., Manola, L., & Huang T. S. (2003). Imagegrouper: a group-oriented user interface for content-based image retrieval and digital image arrangement. *Journal of Visual Languages & Computing* 14 (4), 363-386.

[25] Naphade, M., Smith, J. R., Tesic, J., Chang, S. F., Hsu, W., Kennedy, L., Hauptmann, A., & Curtis, J. (2006). Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13 (3), 86-91.

[26] De Rooij, O., Snoek, C. G. M., & Worring, M. (2008). Mediamill: fast and effective video search using the forkbrowser. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*. *CIVR '08*. (pp. 561-562).

[27] Singla, A., White, R., & Huang, J. (2010). Studying trailfinding algorithms for enhanced web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. *(SIGIR '10)*, 443-450.

[28] Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41 (4), 288-297.

[29] Smeulders, A. W. M., Worring, M. , Santini, S. , A. Gupta, & R. Jain. (2002). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22 (12), 1349-1380.

[30] Snoek, C., Worring, M., Koelma, D., & Smeulders, A. (2006). Learned lexicon-driven interactive video retrieval. Image and Video Retrieval 4071, 11-20.

[31] Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management*, 34 (5), 599-621.

[32] Sun. J.T., Zeng. H.J., Liu. H., Lu. Y., & Chen. Z. (2005). Cubesvd: a novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web*. *WWW '05*. (pp. 382-390).

[33] Urban. J., & Jose, J. M. (2006). Ego: A personalised multimedia management and retrieval tool. *International Journal of Intelligent Systems (Special issue on "Intelligent Multimedia Retrieval")*, 21 (7), 725-745.

[34] Vallet. D., Hopfgartner. F., Jose. J. M., & Castells. P. (2011). Effects of Usage-Based feedback on video retrieval: A Simulation-Based study. *ACM Transaction on Information Systems*, 29(2) 11-32

[35] Villa. R., Gildea. N., & Jose. J. M., A faceted interface for multimedia search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. *SIGIR '08* (pp. 775-776).

[36] White. R. W., Bilenko. M., & Cucerzan. S. (2007) Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '07*. pp. 159-166.

[37] White, R. W. & Huang, J. (2010). Assessing the scenic route: measuring the value of search trails in web logs. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. *SIGIR'10.* (pp. 587-594).

[38] Yang. B., Mei. T., Hua. X.S., Yang. L., Yang. S.-Q., & Li, M. Online video recommendation based on multimodal fusion and relevance feedback. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, ACM Press, 2007, pp. 73-80.

[39] Yang, B., Mei, T., Hua, X.S., Yang, L., Yang, S.Q., & Li, M. (2007). Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval. CIVR '07. (*pp. 73-80).