Toward an Adaptive Video Retrieval System

Frank Hopfgartner and Joemon M. Jose

Department of Computing Science

University of Glasgow

Glasgow, United Kingdom

1. Introduction

With the increasing availability of new tools and applications to record, broadcast and stream videos, there is a need to create new retrieval engines to assist the users in searching and finding scenes they would like to see within different video files. Research to date has a particular emphasis on the system side, resulting in the design of retrieval tools that assist the users in performing search sessions. However, since the effectiveness of current video retrieval systems is everything but satisfying for the users, more sophisticated research is needed to increase the retrieval performance to a similar level as their textual counterparts.

Unlike text retrieval systems, retrieval on digital video libraries is facing a challenging problem: The Semantic Gap. This is the difference between the low-level data representation of videos and the higher level concepts a user associates with video. In 2005, the panel members of the International Workshop on Multimedia Information Retrieval identified this gap as one of the main technical problems in multimedia retrieval (Jaimes et al. 2005), carrying the potential to dominate the research efforts in multimedia retrieval for the next few years. Retrievable information such as textual sources of video clips, i.e. speech transcripts, is often not reliable enough to describe the actual content of a clip. Moreover, the approaches of using visual features

and automatically detecting high level concepts, as mainly studied within the international video processing and evaluation campaign TRECVID (Smeaton et al. 2006), turned out to be not efficient enough to bridge the semantic gap.

One approach to bridge the semantic gap is to improve the interfaces of video retrieval systems, enabling the users to specify their information demand. However, as the performance of state-of-the-art systems indicate, interface designs are, so far, not advanced enough to provide the users with such facilities. A promising approach to solve this problem is to incorporate an adaptive retrieval model, which automatically adapts retrieval results based on the user's preferences. An adaptive retrieval model can be useful to significantly reduce the number of steps the user has to perform before he retrieves satisfying search results. (Sebe and Tian 2007) discuss that for developing an adaptive model for retrieving multimedia content, sophisticated research in various areas is needed, including the acquisition of user preferences and how to filter information by exploiting the user's profile.

(Arezki et al. 2004) provide an example to explain the challenge of different user preferences: When a computer scientist enters the search query "java" into a search engine, he is most likely interested in finding information about the programming language. Other people, however, might expect results referring to the island of Java in Indonesia or a type of coffee beans bearing this name. A classical approach to capture these different preferences is profiling. User profiles can be used to create a simplified model of the user which represents his interests on general topics. Commercial search engines incorporate such profiles, the most prominent being Google offering iGoogle and Yahoo! offering MyYahoo!. Query expansion is used to gather the users' interest and search results are re-ranked to match their interests.

The named services rely on users' explicitly specifying preferences, a common approach in the

text retrieval domain. By giving explicit feedback, users are forced to update their need, which can be problematic when their information need is vague (Spink et al. 1998). Furthermore, users tend to provide not enough feedback on which to base an adaptive retrieval algorithm (Hancock-Beaulieu and Walker 1992).

Deviating from the method of explicitly asking the user to rate the relevance of retrieval results, the use of implicit feedback techniques helps by learning user interests unobtrusively. The main advantage is that users are relieved from providing feedback, it is given unintentionally. An example is printing out a web page, which may indicate an interest in that web page. The basic assumption is that during a search, users' actions are used to maximise the retrieval of relevant information. Implicit indicators have been used and analysed in other domains, such as the WWW (Claypool et al. 2001) and text retrieval (White et al. 2004, Kelly and Teevan 2003), but rarely in the multimedia domain. However, traditional issues of implicit feedback can be addressed in video retrieval since digital video libraries facilitate more interactions and are hence amenable to implicit feedback. (Hopfgartner and Jose 2007) showed that implicit feedback can improve retrieval in digital video library retrieval systems.

A challenging problem in user profiling is the users' evolving focus of interest. What a user finds interesting on day *A* might be completely uninteresting on day *B*, or even on the same day. The following example illustrates the problem: Joe Bloggs is rarely interested in sports. Thus, during Euro 2008, the European Football Championship, he is fascinated by the euphoria exuded by the tournament and follows all reports related to the event. After the cup final, however, his interest abates again. How to capture and represent this dynamic user interest is an unsolved problem. Moreover, a user can be interested in multiple topics, which might evolve over time. Instead of being interested in only one topic at one time, users can search for various independent topics

such as politics or sports, followed by entertainment or business. We can capture this evolution of information need by capturing the implicit factors involved in such a retrieval system.

In this work, we investigate the following research questions: Which implicit feedback a user provides can be considered as a positive indicator of relevance and can hence be used to adapt retrieval results? The second question is how these features have to be weighted to increase retrieval performance. It is not clear which features are stronger and which are weaker indicators of relevance, respectively. Moreover, we aim to study how the users' evolving interest in multiple aspects of news should be considered when capturing the users' interests. Answering these questions will shed light on implicit relevance feedback, a necessary step towards an adaptive retrieval model.

The chapter is organised as follows: A brief introduction of related work is given in Section 2. In Section 3, we discuss research questions which need to be solved in order to develop an adaptive retrieval model. In order to tackle the research questions, we introduce NewsBoy in Section 4, a personalised multimedia application which is designed to capture the user's evolving interest in multiple aspects of news stories. NewsBoy is a web based video retrieval system which enables us to spread the system to a large population, i.e. all students on the University campus. In order to offer an attractive news video retrieval system to the general public, the system is based on an up-to-date news video corpus. NewsBoy automatically processes the daily BBC One news bulletin, segments the broadcast into story segments and recommends news stories by unobtrusively profiling the user based on his interactions with the system. The news aspects are identified by clustering the content of the profile.

2. Background

In the following section, we introduce the multi-faceted research domains which are important in

the scope of this work. In Section 2.1, we provide an overview over the field of interactive video retrieval. Furthermore, we explain the idea of personalised retrieval by incorporating user profiles in Section 2.2 and introduce the users' evolving interest in different aspects of news in Section 2.3. Finally, the idea of feature ranking is introduced in Section 2.4.

In the scope of this research, we aim to rely on both user studies and a simulated user evaluation. Performing user studies is a popular methodology to evaluate interactive retrieval systems. The approach of simulating users to fine tune retrieval systems has been studied before (i.e. Hopfgartner and Jose 2007, Vallet et al. 2008, White et al. 2007), the results being promising to follow the methodology. The evaluation framework is presented in Section 2.5.

2.1 Interactive Video Retrieval Systems

One of the biggest tracks within TRECVID is the interactive search task. In this track, users have to interact with a video retrieval interface in order to retrieve pre-defined topics. Research on interactive video retrieval has been an important stepping stone for the development of large scale video retrieval systems such as YouTube and Google Video. (Snoek et al. 2007) identified an architecture framework for most state-of-the-art video retrieval engines such as (Snoek et al. 2005, Campbell et al. 2006, Rautiainen et al. 2005). This framework can be divided into an indexing engine and a retrieval engine, the first component involving the indexing of the video data. This process starts with a shot segmentation stage, which will split a single video into a sequence of *shots*. A shot is a sequence of the video which is visually related, boundaries between shots typically being marked by a scene cut or fade. Each shot will vary in size, most being very short (typically a few seconds). For each shot, example frames, key frames, are extracted which represent the shot. The shot is used as the element of retrieval: each shot is separately indexed by the system, and the results of searches are presented as a list of shots.

In the news domain, a specific unit of retrieval is the news *story* (Boreczky and Rowe 1996). Examples can be stories about a political event, followed by a story about yesterday's football match or the weather forecast. Therefore, it is necessary to identify and to merge those shots which semantically form one story. However, the structure of news broadcasts directly depends on the programme director's taste, finding a general applicable approach of automatically segmenting a broadcast video into its news stories is hence a challenging task (Chang et al. 2005). With respect to the current systems, indexing shots and stories can be incorporated at a visual, textual and semantic level. (Huang 2003) argues that speech contains most of the semantic information that can be extracted from audio features and according to (Chang et al. 2005); text from speech data has been shown important for key term/named entity extraction, story boundary detection, concept annotation and topic change detection. In literature, the most common text sources are teletext (also called closed-caption), speech recognition transcripts and optical character recognition output. (Hopfgartner 2007) compares different state-of-the-art video retrieval systems, concluding that the text sources for systems differ significantly from each other. While (Heesch et al. 2004) include closed caption transcripts, automatic speech recognition output and optical character recognition output into their index, whereas (Foley et al. 2005) index speech recognition output only.

The indexing and the related retrieval methods make up the "backend". The second component, the "frontend", is the interface between the computer and the human user. Graphical user interfaces give the user the opportunity to compose queries with the retrieval engine handling these queries, combining returned results and visualising them. A detailed survey of different interface approaches is provided by (Hopfgartner 2007). In the remainder of this section, we focus on systems which incorporated the idea of providing users with a news video retrieval

system that is based on daily news videos.

(Pickering et al. 2003) record the daily BBC news and capture the broadcasted subtitles. The news broadcast is segmented into shots and key frames extracted to represent the shot. Shots are merged to form news stories based on the subtitles. Therefore, they extract key entities (nouns, verbs, etc.) from the subtitles and calculate a term weighting based on their appearance in the teletext. The interface of their system is web based[1] and provides browsing and retrieval facilities. They concentrated their work on summarising news video, adapting retrieval results to the user's need has not been considered in their system.

(Morrison and Jose 2004) introduce the web based news video retrieval system VideoSqueak. They record the BBC One evening news and use the captured subtitles to identify story units. The subtitles of the broadcast are the retrieval source. They evaluate different presentation strategies for multimedia retrieval. However, they have not studied the user behaviour in news retrieval systems.

2.2 Personalisation

Web 2.0 facilities enable everyone to easily create their own content and to publish it online. Users can upload videos on platforms such as YouTube, share pictures on Flickr or publish anything in a weblog. Two direct consequences of this development can be identified: First of all, it leads to a growing quantity of content presented in a multimedia format. Secondly, information sources are completely unstructured and finding interesting content can be an overwhelming task. Hence, there is a need to understand the user's interest and to customise information accordingly.

---

[1] A demo can be found online at http://www.doc.ic.ac.uk/~mjp3/anses

A common approach to capture and to represent these interests is user profiling. Using user profiles to create personalised online newspapers has been studied for a long time.

(Chen and Sycara 1998) join internet users during their information seeking task and explicitly ask them to judge the relevance of the pages they visit. Exploiting the created user profile of interest, they generate a personalised newspaper containing daily news. However, providing explicit relevance feedback is a demanding task and users tend not to provide much feedback (Hancock-Beaulieu and Walker 1992).

(Bharat et al. 1998) create a personalised online newspaper by unobtrusively observing the user's web-browsing behaviour. Although their system is a promising approach to release the user from providing feedback, their main research focus is on developing user interface aspects, ignoring the sophisticated retrieval issues.

(Smeaton et al. 2002) introduced Físchlár-News, a news video recommendation system that captured the daily evening news from the national broadcaster's main TV channel. The web based interface of their system provides a facility to retrieve news stories and recommends stories to the user based on his interest. According to (Lee et al. 2006), the recommendation of Físchlár-News is based on personal and collaborative explicit relevance feedback. The use of implicit relevance feedback as input has not been incorporated. Profiling and capturing the users is an important step towards adapting systems to the user's evolving information interest. In the following section, we introduce the ostensive model which allows us to capture users' evolving interest.

2.3 Evolving User Interest

In a retrieval context, profiles can be used to contextualise the user's search queries within their interests and to re-rank retrieval results. This approach is based on the assumption that the user's

information interest is static, which is however, not appropriate in a retrieval context. (Campbell 1995) argues that the users' information need can change within different retrieval sessions and sometimes even within the same session. He states that the user's search direction is directly influenced by the documents retrieved. The following example explains this observation: Imagine a user who is interested in red cars and uses an image retrieval system to find pictures showing such cars. His first search query returns him several images including pictures of red Ferraris. Looking at these pictures, he wants to find more Ferraris and adapts the search query accordingly. The new result list now consists of pictures showing red and green Ferraris. Fascinated by the rare colour for this type of car, he again re-formulates the search query to find more green Ferraris. Within one session, the user's information need evolved from red cars to green Ferraris. Based on this observation, (Campbell and van Rijsbergen 1996) introduce the ostensive model which incorporates this change of interest by considering when a user provided relevance feedback. In the ostensive model, providing feedback on a document is seen as ostensive evidence that this document is relevant for the user's current interest. The combination of this feedback over several search iterations provides ostensive evidence about the user's changing interest. The model considers the user's changing focus of interest by granting the most recent feedback a higher impact over the combined evidence. Various forms of this model have been developed and applied in image retrieval (Urban et al. 2003) and web search scenarios (Joho et al. 2007).

In this section, we discussed the challenge of capturing the evolving interest of the users based on interpreting their interactions with a system. In the next section, we discuss the problem of ranking this feedback in order to adapt a retrieval model to the users' need.

2.4 Relevance Ranking

Most information retrieval systems such as Google or Yahoo! attempt to rank documents in increasing order of relevance. A major challenge in the field is how to judge whether a document is relevant to a given query or otherwise. One approach is to rank results based on query dependent features such as the term frequency of a document and the distribution of each term in the entire collection. Hence for each query, results are ranked based on a dynamic relevance score. However, better retrieval performance can be achieved by transforming a document's query independent features into a static relevance score and including this in the overall relevance score. Thinking of a textual corpus such as the WWW, query independent features can be the amount of hyperlinks that point to a document or the length or creation time of a document. Here, the major challenge is combining scores and to define good functions that transform the documents into applicable weighting schemes. (Craswell et al. 2005) explain that a retrieval model which incorporates these query independent features has to answer basic questions. The first question is whether the feature is needed to adjust the weighting of a document. It might not be necessary to incorporate the weighting of a specific feature if the initial weighting is already appropriate. The second question is how different features shall alter the weighting of a document. And finally, the model should predict that the created weighting represents best the document.

According to (Craswell et al. 2005), various approaches have been studied to combine different features, rank-based, and language modelling priors being most promising. (Fagin et al. 2003) and (Cai et al. 2005) combine features by creating document ranking lists for each feature and merging these lists based on their score. The advantage of this approach is that the diversified weighting used in the different feature ranking lists does not matter, as the combined list takes

only the score within each list into account.

(Kraaij et al. 2002) calculate prior probabilities for various features such as page length and URL type to increase the precision of a text retrieval system. Even though they conclude that using priors to combine independent features can improve the retrieval performance, they argue that choosing a wrong prior can decrease the performance. So it is important to identify when a prior is useful.

Given that we provided a survey over related work of our research domain, we will introduce different evaluation methodologies which we aim to use in order to study our research questions in the following section.

2.5 Evaluation Framework

A common approach to study the users' behaviour of interacting with a computer system is to perform a user study, to monitor the users' interactions and to analyse the resulting log files. Such an analysis shall help to identify good implicit indicators of relevance, as it can help to answer basic questions: What did the user do to find the information he/she wanted? Can the user behaviour be used to improve retrieval results?

To get an adequate impression of users' behaviour when interacting with a video retrieval system, we need a large quantity of different users interacting with the system which is necessary to draw general conclusions from the study, i.e. by analysing user log files. Besides, non-expert users should be interacting with the system, as they will interact in a more intuitive way than expert users. However, it is not practical to conduct such a study in all situations, mainly due to cost associated with them. Besides, it is hardly possible to benchmark different parameter combinations of features for effectiveness using user-centred evaluations.

An alternative way of evaluating such user feedback is the use of simulated interactions. In such

approach, a set of possible steps are assumed when a user is performing a given task with the evaluated system. (Finin 1989) introduced one of the first user simulation modelling approaches. This "General User Modelling System" (GUMS) allowed software developers to test their systems in feeding them with simple stereotype user behaviour. (White et al. 2007) proposed a simulation-based approach to evaluate the performance of implicit indicators in textual retrieval. They simulated user actions as viewing relevant documents, which were expected to improve the retrieval effectiveness. In the simulation-based evaluation methodology, actions that a real user may take are assumed and used to influence further retrieval results. (Hopfgartner et al. 2007) introduced a simulation framework to evaluate adaptive multimedia retrieval systems. In order to develop a retrieval method, they employed a simulated evaluation methodology which simulated users giving implicit relevance feedback. (Hopfgartner and Jose 2007) extended this simulation framework and simulated users interacting with state-of-the-art video retrieval systems. They argue that a simulation can be seen as a pre-implementation method which will give further opportunity to develop appropriate systems and subsequent user-centred evaluations. (Vallet et al. 2008) use the concept of simulated actions and try to mimic the interaction of past users by simulating user actions based on the past history and behaviour of users with an interactive video retrieval system. Their study has proven to facilitate the analysis of the diverse types of implicit actions that a video retrieval system can provide.

Analysing these research efforts lead to the conclusion that even though simulation based studies should be confirmed by user studies, they can be a cheap and repeatable methodology to fine tune video retrieval systems. Hence, user simulation is a promising approach to further study adaptive video retrieval, at least as a preliminary step.

3. Research Framework

The scope of this research is to develop an adaptive video retrieval model, which automatically adapts retrieval results to the users information need. In the previous section, we therefore introduced various aspects which are relevant within this scope, including an introduction to interactive video retrieval, personalisation approaches, the users' evolving interests, ranking approaches and different evaluation methodologies. In this section, we focus on research questions which need to be tackled in order to develop an adaptive video retrieval system. More particular, we are interested in the following problems:

1. Which implicit feedback a user provides while interacting with an interface can be considered as a positive indicator of relevance?
2. Which interface features are stronger indicators of relevance, or: How shall these features be weighted in order to increase retrieval performance?
3. How should the user's evolving interest in multiple aspects of news be incorporated when retrieval results are re-ranked in accordance to his interest?

Once the users' intentions and information need is clear, systems can be built that take advantage of such knowledge and optimise the retrieval output for each user by implementing an adaptive video retrieval model.

In order to study the first research question, we will provide users with different video retrieval interface approaches for different interaction environments such as desktop PCs of iTV boxes. Hence, users are required to interact differently with the interfaces. The difference has a strong influence on the user's behaviour, making the importance of implicit indicators of relevance application-dependent. Comparing user interactions with different applications should help to identify common positive indicators. The research will be conducted around two different

applications where we can monitor user feedback: desktop computers and television. The specific characters of these environments are introduced in the following.

- **Desktop Computers:** The most familiar environment for the user to do video retrieval is probably a standard desktop computer. Most adaptive video retrieval systems have been designed to run under such environment. The interface can be displayed on the screen and users can easily interact with the system in using the keyboard or mouse. One can assume that users will take advantage of this interaction and hence give a high quantity of implicit feedback. From today's point of view, this environment offers the highest amount of possible implicit relevance feedback. An example interface is introduced in Section 4.2.
- **iTV:** A widely accepted medium for multimedia consumption is the television. Watching television, however, is a passive procedure. Viewers can select a programme using a remote control, changing the content is not possible though. Recently, Interactive TV is becoming more and more popular. Using a remote control, viewers can interact directly when watching television, e.g. in participating in quiz shows. In news video retrieval, this limited interaction is a challenge. It will be more complex to enter query terms, e.g. in using the channel selection buttons as it is common for remote controls. Hence, users will possibly avoid entering key words. On the other hand, the selection keys and a display on the remote control provide a method to give explicit relevance feedback. An example: The viewer sees a video segment on television. Now, he/she uses the remote control to judge the relevance of this segment.

A well studied research methodology in the information retrieval community to evaluate different parameters or environments is to perform user studies. Analysing users' interactions

with different interface approaches will help us to understand how users interacted with this application, and will lead to further knowledge which interface features are general indicators of relevance. Furthermore, we will use the simulation methodology introduced in Section 2.5, i.e. by exploiting the user log files as applied by (Hopfgartner et al. 2008) and analysing the effect of different feature weighting schemes on retrieval performance. This analysis should help to distinguish stronger and weaker indicators of relevance and hence will answer our second research question.

Moreover, we aim to study the users' evolving interest and different ranking approaches in order to answer the third introduced research question. In the following section, we introduce NewsBoy, the system we aim to use to further investigate the above introduced research questions. Our first choice is to rely on the previously introduced ostensive model. NewsBoy introduces an example interface designed for the use on desktop computers and incorporates a simple approach of capturing the users' evolving interest based on the ostensive model.

4. NewsBoy Architecture

In order to study our research questions, we need many users interacting with a video retrieval system. Hence, we developed NewsBoy, a web based news video retrieval system based on AJAX technology for personalised news retrieval. This will enable us to spread the system to a large population, i.e. all students on a University campus. AJAX takes away the burden of installing additional software on each client (assuming that JavaScript is activated and a Flash Player running on the client side). Besides, due to the popularity of Web 2.0 technology, users get used to interact with complex applications using their browser only. This might motivate them to use the system on a regular basis to retrieve broadcasting news.

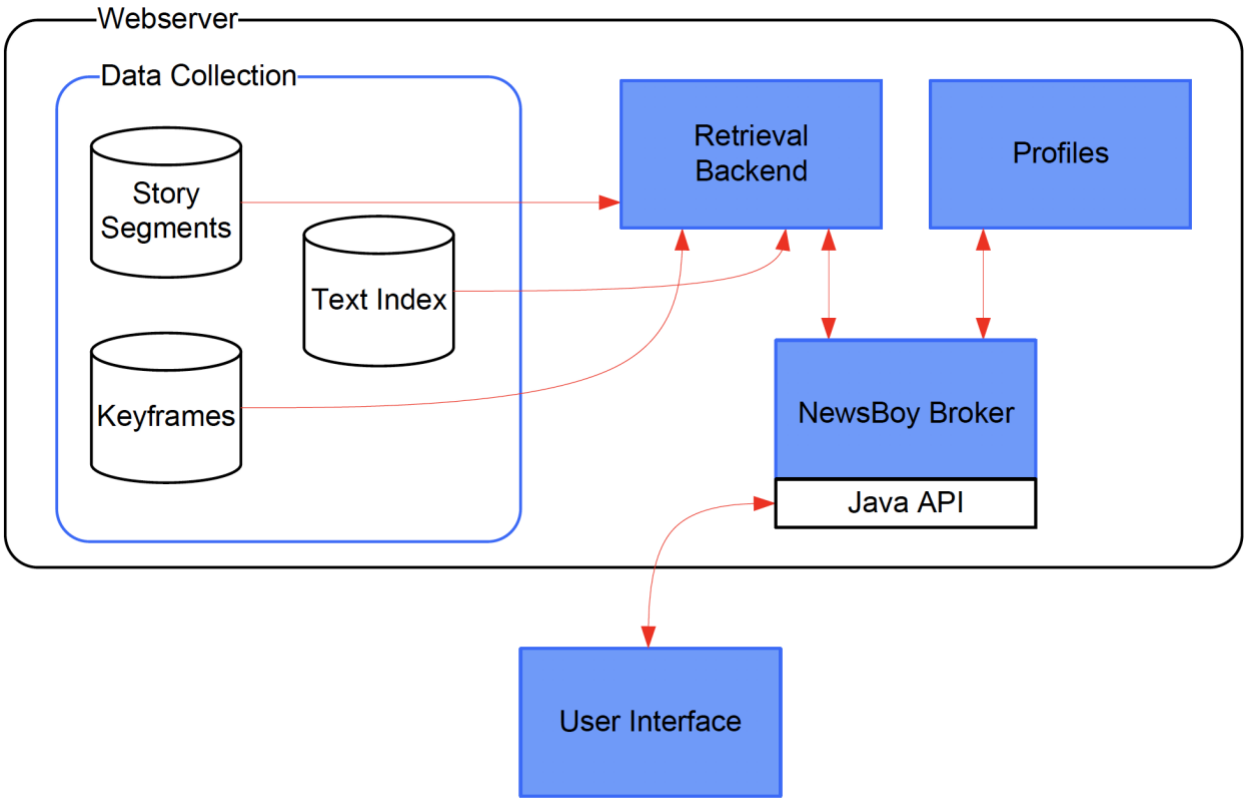

Figure 1. NewsBoy Architecture

Figure 1 illustrates the conceptual design of the system. As the graphic shows, NewsBoy can be divided into five main components, four running on a web server and one, the user interface, on the client side. The first component is the data collection. The process of recording, analysing, and indexing a daily news broadcast to create a data collection will be introduced in Section 4.1. The retrieval backend, the second component of NewsBoy, administers the data collection. We are using MG4J[1], an open source full-text search engine. As argued in Section 3, we aim to provide users with different interface approaches for different interaction environments such as desktop PCs of iTV boxes. One interface approach designed for the use on desktop PCs, the third

[1] http://mg4j.dsi.unimi.it/

component of the NewsBoy architecture will be introduced in Section 4.2. In Section 4.3, we introduce the concept of capturing the users' interests in user profiles by interpreting their interactions with these interfaces. It is the fourth component of the NewsBoy system and aims to answer the previously introduced third research question by incorporating the ostensive model.

4.1 Data Collection

In recent years, the retrieval of news video data has been the main focus of research in the field of interactive video retrieval. A main reason for this concentration on this domain is the international TRECVID (Smeaton et al. 2006) workshop, which provided a large corpus of news videos in the last few years. However, the latest data collection consisting of news video was recorded in late 2005. While these videos can be used to measure the system-centred research approaches, it is not recommendable to base long term user studies on this old data set. Most likely, potential users will get bored with the outdated data which eventually results in a lack of motivation to search for interesting topics within the corpus and hence biases the study. One method to avoid this effect is to provide users with up-to-date news videos.

In this section, we describe the process of recording a daily news bulletin and introduce our approach of segmenting the broadcast into news stories, the unit of retrieval in our system. We focus on the regional version of the BBC One O'Clock news. The programme covers international, national (UK) and regional (Scotland) topics, which are usually presented by a single newsreader. The BBC enriches its television broadcast with Ceefax, a closed caption (teletext) signal which provides televisual subtitles. The bulletin has a running time of 30 minutes and is broadcasted every day from Monday till Friday on BBC One, the nation's main broadcasting station. Both analogue and digital broadcasts can be received by aerial antennas. In addition, the BBC streams the latest programme on their website.

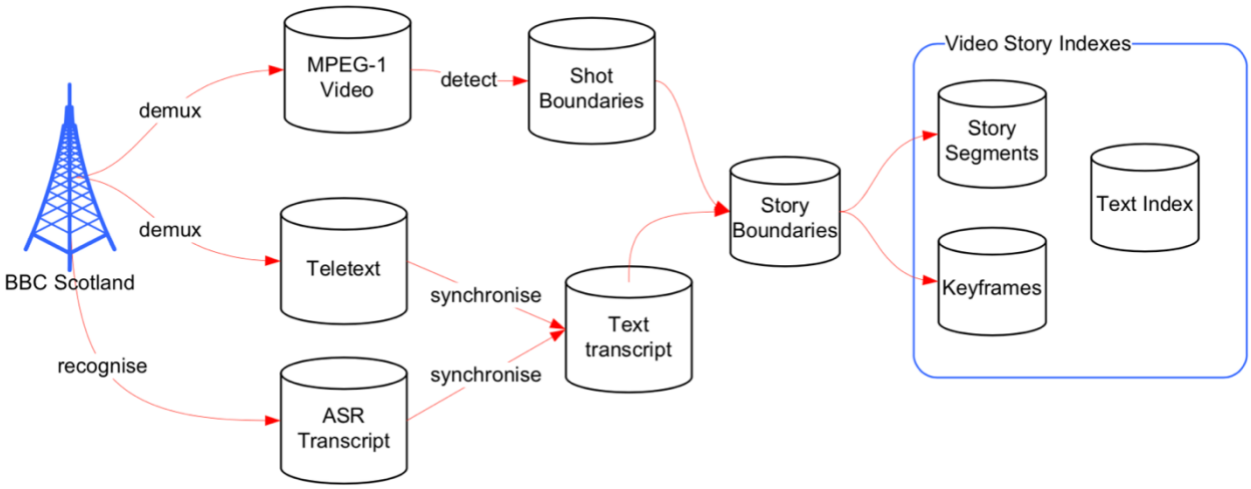

Figure 2. System architecture of the capturing and indexing process

Figure 2 illustrates the architecture for recording, analysing and indexing the news programme broadcasted by BBC Scotland. The process is divided into a shot boundary detection and key frame extraction task, followed by the creation of a text transcript. Both shot boundaries and text transcript are used to identify story units. The different steps will be introduced in the remainder of this section.

The video/audio stream is downloaded from the BBC website, where it is provided in Windows Media Format as an online stream. Adopting the techniques introduced by (O'Toole et al 1999, Browne et al. 2000), we use a colour histogram-based method to detect shot boundaries in our video files. Furthermore, we detect example key frames by calculating the average colour histogram for each shot and extract the frames within the shot which are closest to the average. In an additional step, we combine the key frames belonging to the same shot to form an animated presentation of the shot.

Further, we capture the teletext by decoding the aerial transmission signal of the BBC. The BBC's Ceefax system was developed to provide televisual subtitles for the deaf. They are manually created, thus the semantic quality of the text is reliable. However, the text is not

synchronised with the actual speech. According to (Huang 2003), the mean delay between speech and teletext is between 1 to 1.5 seconds. Furthermore, the transcript does not always represent the whole spoken text, but more likely a shortened version of the sentences. While these two drawbacks are acceptable when the text is considered as an additional service to accompany the news programme, it can be problematic when used as the source of a content analysis. Therefore, we create a second text transcript by performing an automatic speech recognition using the Sphinx III[1] system. Sphinx III is a speaker independent speech recognition system which, according to (Huang 2003), is the best performing tool for news video data sets. For recognition, we use the open source acoustic models, language models and dictionaries provided under the Open Source licence. As these are US English models, the recognition of the BBC news broadcast, mainly spoken in British English, is rather weak. This means that the teletext transcript contains more correct words while the transcript provides correct time codes for positively identified terms. It is therefore necessary to improve the text transcript. Following (Huang 2003), we merge both closed caption and ASR transcripts by aligning closed caption terms with time codes and synchronise these terms with the time codes of the ASR transcript. Within a sliding window of plus/minus 2.5 seconds around the time code of a term appearing in the teletext list, we calculate the Levenshtein distance for each word. If the distance stays below a predefined threshold, we merge the ASR transcript's time code with the word found in the teletext list. Terms are considered to be actual spoken terms when they appear in both streams within the defined time window. However, in several cases, no matching term is available within the time window. In this case, we assume that the quality of the speech recognition output is too bad and hence use the teletext term and its output only.

---

[1] http://cmusphinx.sourceforge.net/

As most stories on the BBC One O'Clock news are introduced by the anchorman or a speaker in the background, we divide the audio stream into speaker segments using the free available tool mClust[1]. We then analyse the text transcript of these speaker segments to identify whether segments can be merged to a story segment candidate. In a first step, we merge "interview" segments that are segments spoken by the same speaker which are interrupted by another, short, speaker segment.

Figure 3. Speaker Segments

Figure 3 illustrates an example. Here, three different speakers have been identified: S_1, S_2 and S_3. The speaker segment S_2 is surrounded by two segments of speaker S_1. Assuming that speaker S_1 is a journalist who interviews speaker S_2 and afterwards continues with the story, we merge the three mentioned segments to one story segment candidate. In a next step, we scan the segments for *signal terms* such as "Welcome to…", "Thank you…" or "Let's speak to…" which indicate the end or beginning of a story. Further, we use the Spearman rank-order correlation to compute the degree of similarity between neighboured segments. The correlation returns values between −1 and 1, where 0 shows that there is no correlation between the segments and 1 shows that they are a perfect match. (White et al. 2003) showed the use of a similarity threshold of 0.2 in a text retrieval scenario which we found useful in this case. Moreover, we match the detected story unit

[1] http://www-lium.univ-lemans.fr/tools

candidates with the detected shot boundaries, assuming that a news story always begins with a new shot. To further enrich the segmented stories, we use the General Architecture for Text Engineering (GATE)[1] to identify persons, locations and relative time mentioned in the transcript.

4.2 Desktop PC Interface

In this section, we present an example interface as it can be used in a desktop PC environment. It provides various possibilities to provide implicit relevance feedback. Users interacting with it can:

- Expand the retrieved results by clicking on it.
- Play the video of a retrieved story by clicking on “play video”.
- Play the video for a certain amount of time.
- Browse through the key frames.
- Highlight additional information by moving the mouse over the key frames.

[1] http://gate.ac.uk/

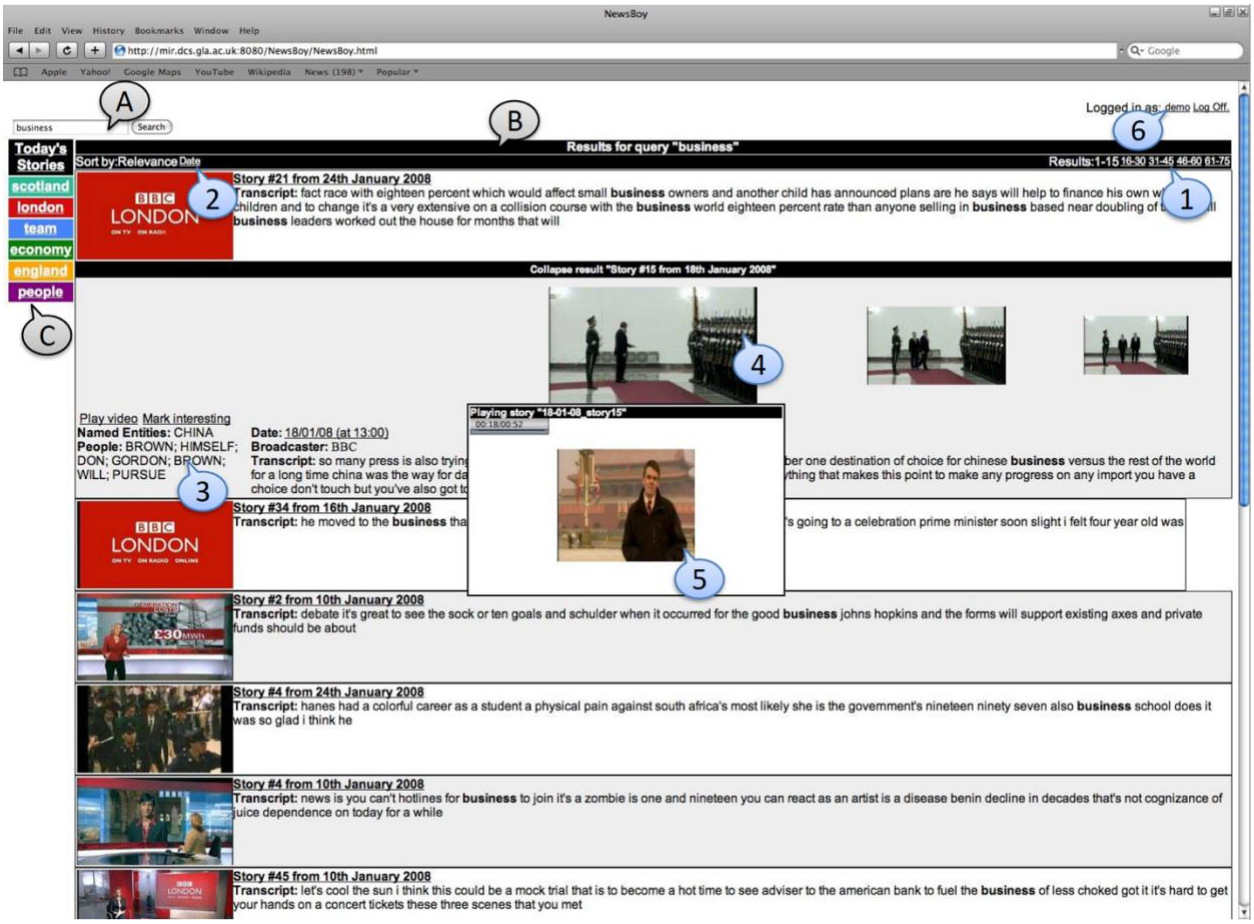

Figure 4. Example of a desktop PC environment interface

Figure 4 shows a screenshot of the interface, its features will be described in the following section. The interface can be divided into three main panels, search panel (A), result panel (B) and clustered search queries (C).

In the search panel (A) users can formulate and carry out their searches by entering a search query and clicking the button to start the search. BM25 (Robertson et al. 1994) is used to rank the retrieved documents in accordance to their relevance to a given search query.

Once a user logs in, NewsBoy displays the latest news stories in the result panel (B). Furthermore, this panel lists retrieval results. The panel displays a maximum of 15 results; further results can be displayed by clicking the annotated page number (1). The results can be sorted in accordance to their relevance to the query or chronologically by their broadcasting date

(2). Results are presented by one key frame and a shortened part of the text transcript. A user can get additional information about the result by clicking on either the text or the key frame. This will expand the result and present additional information including the full text transcript, broadcasting date, time, channel and a list of extracted named entities such as persons, locations and relative times (3). In the example screenshot, the second search result has been expanded. The shots forming the news story are represented by animated key frames of each shot. Users can browse through these animations by clicking on the key frame. This action will centre the selected key frame and surround it by its neighboured key frames. The key frames are displayed in a fish-eye view (4), meaning that the size of the key frame grows larger the closer it is to the focused key frame. In the expanded display, a user can also select to play a video or to mark it as interesting. Clicking on "play video" starts playing the story video in a new panel (5).

NewsBoy recommends daily news videos based on the user's multi-aspect preferences. These preferences are captured by unobtrusively observing the user's interactions with the NewsBoy interface. By clustering the content of the profile, NewsBoy identifies different topics of interest and recommends these topics to the user. The interface presents these topics as labelled clusters on the left hand side of the interface (C). Each cluster represents a group of terms, hence, when a user clicks on the term, a new search is triggered, using the selected terms as a new query. Results are displayed in the result panel.

On the top of the interface, the users can edit their profile by clicking on their username (6). This action will pop up a new frame where the top weighted terms of each cluster are listed, and the user can edit terms or the aligned weighting. Furthermore, the user can manually add new weighted terms.

4.3 Profile

User profiling is the process of learning the user's interest over a longer period of time. Several approaches have been studied to capture a user's interest in a profile, the most prominent being the weighted keyword vector approach. In this section, we introduce the approach and introduce the problems which occur in capturing this interest in a profile.

In the weighted term approach, interests are represented as a vector of weighted terms where each dimension of the vector space represents a term aligned with a weighting. Hence, in order to capture the user's interest, terms aligned with the story item a user interacted with should be extracted and weighted with the feedback based on the user's interaction. The weighting of the terms will be updated when the system submits a new set of weighted terms to the profile starting a new iteration j. Hence, the interaction I of a user i at iteration j can be represented as a vector of weights

$$\overrightarrow{\mathrm{I_{ji}}} = \{\mathrm{W_{ij1}} \ldots W_{\mathrm{ijv}}\}$$

where v indexes the word in the whole vocabulary V. The weighting W_{ij} depends on the implicit relevance feedback provided by a user i in the iteration j while interacting with an interface. Identifying an optimal weighting for each interface feature W_{ij} is one of the research questions we aim to study. Once the weighting has been determined, representative terms from relevant documents will be extracted and assigned with an indicative weight to each term, which represents its weight in the term space. In a simple model we propose, we extract non-stopwords v from the stories a user interacted with the iteration i and assign these terms with the relevance weighting $\mathrm{W_{ijv}}$. Furthermore, the profile $\overrightarrow{P_{\mathrm{i}}}$ of user i can be presented as a vector containing the profile weight PW of each term v of the vocabulary:

$$\vec{P_i} = \{PW_{i1} \ldots PW_{iv}\}$$

4.3.1 Capturing Evolving Interest

The simplest approach to create a weighting for each term in the profile is to combine the weighting of the terms over all iterations. This approach is based on the assumption that the user's information interest is static, which is, however, not appropriate in a retrieval context. The users' information need can change within different retrieval sessions and we aim to study how this change of interest can be incorporated.

(Campbell and van Rijsbergen 1996) propose in their ostensive model that the time factor has to be taken into account, i.e. by modifying the weighting of terms based on the iteration they were added to the user profile. They argue that more recent feedback is a stronger indicator of the user's interest than older feedback. In our profile, the profile weight for each user i is the combination of the weighted terms v over different iterations j: $PW_{\text{iv}} = \sum_j a_{\text{j}} W_{\text{ijv}}$. We include the ostensive factor, denoted a_j, to introduce different weighting schemes based on the ostensive model. We have experimented with four different functions to calculate the weighting, depending on the nature of aging:

- Constant weighting
- Exponential weighting
- Linear weighting
- Inverse exponential weighting

Results of a user-centred evaluation of these weighting functions are discussed in (Hopfgartner et al. 2008b). Figure 5 plots the normalised functions for up to ten iterations. It can be seen that all functions, apart from the constant weighting, reduce the ostensive weighting of earlier iterations. The weighting depends on the constant $C > 1$. The functions will be introduced in the

remainder of this section.

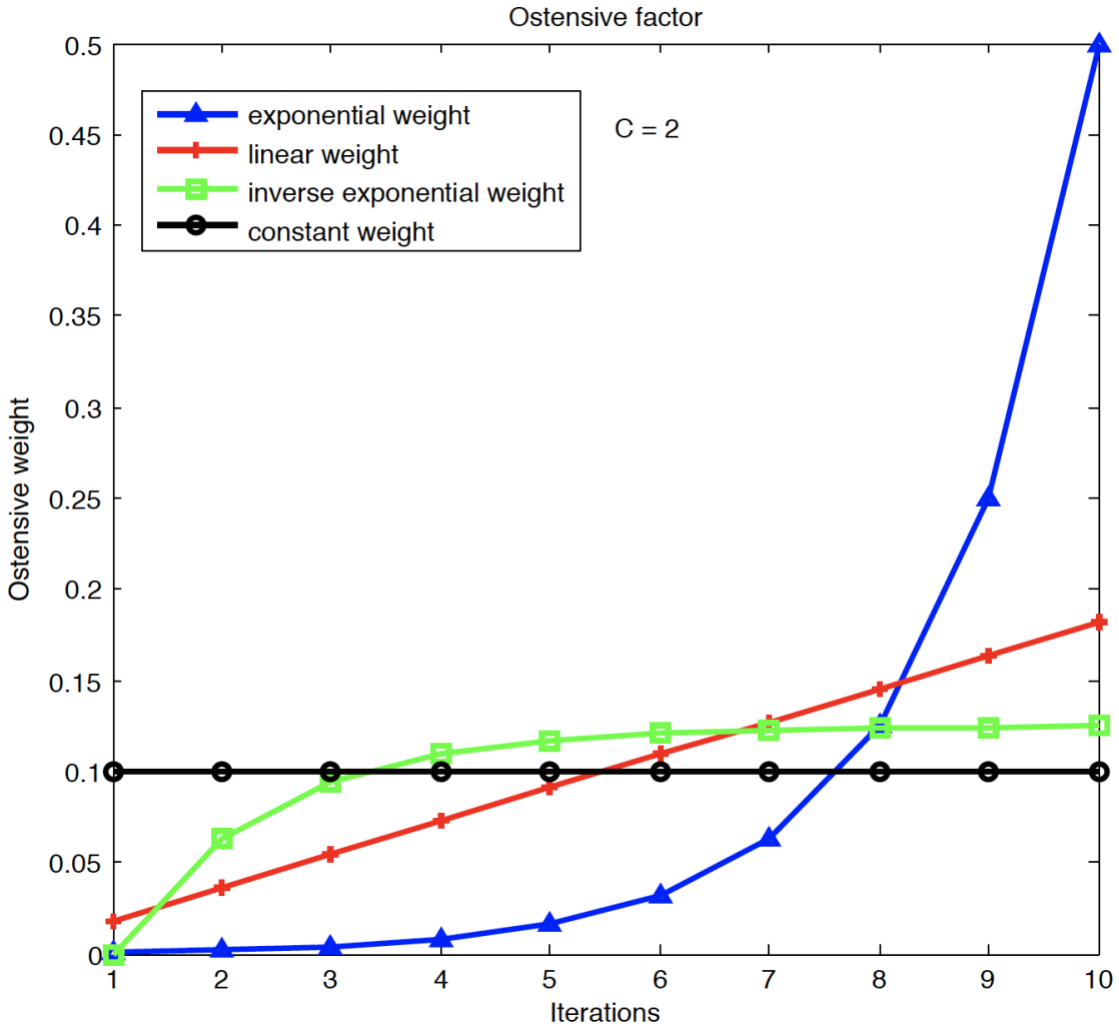

Figure 5. Ostensive weighting over ten iterations

Constant Weighting

Equation 1

$$a_j = \frac{1}{j_{max}}$$

The constant weighting function does not influence the ostensive weighting. As Equation 1 illustrates, all terms will be combined equally, ignoring the iteration when a term was added or updated. The constant weighting can be seen as a baseline methodology which does not include

any ostensive factor.

Exponential Weighting

Equation 2

$$a_{\mathrm{j}} = \frac{C^{j}}{\sum_{k=1}^{j_{max}} C^{k}}$$

The exponential weighting as defined in Equation 2 gives a higher ostensive weighting to terms which has been added or updated in older iterations. It is the most extreme function as the ostensive weighting of earlier iterations decreases distinctly.

Linear Weighting

Equation 3

$$a_{\mathrm{j}} = \frac{Cj}{\sum_{k=1}^{j_{max}} Ck}$$

Equation 3 defines the linear weighting function. The ostensive weighting of earlier iterations decreases linearly. This function linearly reduces the ostensive weighting of earlier iterations.

Inverse Exponential Weighting

Equation 4

$$a_{\mathrm{j}} = \frac{1 - C^{-j+1}}{\sum_{k=1}^{j_{max}} 1 - C^{-k+1}}$$

The inverse exponential weighting defined by Equation 4 is the most contained function. Compared to the other introduced functions, the ostensive weighting of early iterations decreases more slowly.

4.3.2 Capturing Multiple Interests

All components introduced in the previous sections communicate through the NewsBoy Broker, the fifth component of the system illustrated in Figure 4. The task of the broker is to personalise the system by identifying the user's multiple interests in different aspects. Our methodology of

identifying these aspects, the third research question as introduced in Section 3, is introduced in the following.

We base our approach on the assumption that news topics consist of a number of particular terms which appear in all stories about this topic. News stories about the topic football e.g. might consist of unique terms such as “goal”, “offside”, “match” or “referee”. We capture implicit feedback when a user interacts with these stories. The terms of these stories will be extracted and, combined with the implicit weighting, stored in the profile. Hence, as the particular terms are added with the same weighting, they are close neighbours in the profile’s vector space.

In this work, we limited the number of different aspects to a maximum of six. Therefore, we sort the terms in the user’s profile according to their profile weighting and identify the terms which have the five biggest distances to the neighboured terms. We use these identified weighted terms to cluster the remaining profile terms accordingly. Each cluster represents one aspect of the user’s interest.

The top weighted terms of each cluster are used as a label to visualise the aspect on the left hand side of the NewsBoy interface (marked (C) in Figure 4). Clicking on this label hence triggers a retrieval with the top six weighted terms of this aspect being used as search query.

5. Discussion

When comparing video and text retrieval systems, one notices a large difference in retrieval performance. State-of-the-art systems are not yet advanced enough to understand the user’s interest and to identify relevant video scenes. The Semantic Gap has been identified to be the main reason for this problem. While humans can easily understand the content of images or videos, computers are not capable of doing so. Different approaches are currently studied to bridge this gap, the most prominent being the automatic detection of high level concepts in a

video. However, this approach has not been efficient and effective enough. A second approach is to improve the query formulation schemes so that a user can accurately specify queries. However, as the performance of state-of-the-art systems indicate, interface designs are not advanced enough to provide the users with facilities to enter their information need. Hence, we argue that there is a need for more sophisticated interfaces to search for videos.

In this work, we propose to adapt retrieval based on the user's interaction with video retrieval interfaces. In the text retrieval domain, the approach of interpreting the user's action as implicit indicator of relevance turned out to be an effective method to increase retrieval performance. In the video retrieval domain, however, rarely anything is known about which implicit feedback can be used as implicit indicators of relevance. We focus on three questions: The first problem we discussed is which implicit feedback a user provides can be considered as a positive indicator of relevance and can hence be used to adapt retrieval results. The second problem is how these features have to be weighted in order to increase retrieval performance. It is not clear which features are stronger and which are weaker indicators of relevance, respectively. Moreover, we argued that the users' evolving interest in multiple news aspects has to be considered when capturing the users' interests.

We discussed different evaluation approaches to tackle this research problem, including performing a user study in order to analyse the user's interactions with a video retrieval system and simulating users by exploiting log files of the users' interactions with the system. As a basis of our research, we introduced NewsBoy, a web based news video retrieval system. NewsBoy captures and processes the daily BBC One O'Clock news bulletin and provides an interface to access this data. The introduced system can be seen as a medium which will be used to answer the introduced research questions.

Acknowledgement

This research was supported by the European Commission under the contract FP6-027026-K-SPACE.

References

Arezki, R., Poncelet, P., Dray, G., and Pearson, D. W. (2004). Adaptive Hypermedia and Adaptive Web-Based Systems, *Web Information Retrieval Based on User Profiles*, pp. 275–278. Springer Verlag.

Bharat, K., Kamba, T., and Albers, M. (1998). Personalized, Interactive News on the Web. *Multimedia Systems* 6(5), 349–358.

Boreczky, J. S. and Rowe, L. A. (1996). Comparison of Video Shot Boundary Detection Techniques. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 170–179.

Browne, P., Smeaton, A. F., Murphy, N., O'Connor, N., Marlow, S., and Berrut, C. (2000). Evaluating and Combining Digital Video Shot Boundary Detection Algorithms. In *IMVIP 2000: Proceedings of the Irish Machine Vision and Image Processing Conference.* Belfast, Northern Ireland.

Cai, D., He, X., Wen, J.-R, and Ma, W.-Y. (2004). Block-level link analysis. In *SIGIR '04: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pp. 440–447. ACM Press.

Campbell, I. (1995). Supporting Information Needs by Ostensive Definition in an Adaptive Information Space. In *MIRO'95: Workshops in Computing*. Springer Verlag.

Campbell, I. and van Rijsbergen, C. J. (1996). The Ostensive Model of Developing Information

Needs. In *Proc. of CoLIS-96, 2nd Int. Conf. on Conceptions of Library Science*, pp. 251–268.

Campbell, M., Haubold, A., Ebadollahi, S., Naphade, M. R., Natsev, A., Seidl, J., Smith, J. R., Tešić, J., and Xie, L. (2006). IBM Research TRECVID-2006 Video Retrieval System. In *TRECVID 2006: Text Retrieval Conference, TRECVID Workshop,* Gaithersburg, Maryland, November 2006.

Chang, S.-F., Manmatha, R., and Chua, T.-S. (2005, 03). Combining Text and Audio-Visual Features in Video Indexing. In *ICASSP'05 – Proceedings of Acoustics, Speech, and Signal Processing Conference*, pp. 1005–1008.

Chen, L. and Sycara, K. (1998). WebMate: A Personal Agent for Browsing and Searching. In K. P. Sycara and M. Wooldridge (Eds.), *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, New York, pp. 132–139. ACM Press.

Claypool, M., Le, P., Wased, M., and Brown, D. (2001). Implicit Interest Indicators. In *Intelligent User Interfaces*, pp. 33–40.

Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. (2005). Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28thAnnual International ACM SIGIR conference on Research and Development in Information Retrieval*, New York, NY, USA, pp. 416–423. ACM Press.

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A., and Williamson, D. P. (2003). Searching the workplace web. In *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, New York, NY, USA, pp. 366–375. ACM Press.

Finin, T. W. (1989). GUMS: A General User Modeling Shell. *User Models in Dialog Systems*,

411–430.

Foley, E., Gurrin, C., Jones, G., Lee, H., McGivney, S., O'Connor, N. E., Sav, S., Smeaton, A. F., and Wilkins, P. (2005). TRECVid 2005 Experiments at Dublin City University. In *TRECVid 2005: Text REtrieval Conference, TRECVID Workshop*, Gaithersburg, Maryland, 14-15 November 2005.

Hancock-Beaulieu, M. and Walker, S. (1992). An Evaluation of Automatic Query Expansion in an Online Library Catalogue. *J. Doc.* 48(4), 406–421.

Heesch, D., Howarth, P., Magalhães, J., May, A., Pickering, M., Yavlinski, A., and Rüger, S. (2004). Video Retrieval using Search and Browsing. In *TREC2004: Text REtrieval Conference*, Gaithersburg, Maryland, 15-19 November 2004.

Hopfgartner, F. (2007). *Understanding Video Retrieval*. Saarbrücken, Germany: VDM Verlag.

Hopfgartner, F. and Jose, J. (2007). Evaluating the Implicit Feedback Models for Adaptive Video Retrieval. In *ACM MIR '07: Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Augsburg, Germany, pp. 323–332.

Hopfgartner, F., Urban, J., Villa, R., and Jose, J. (2007). Simulated Testing of an Adaptive Multimedia Information Retrieval System. In *CBMI'07: Proceedings of the Fifth International Workshop on Content-Based Multimedia Indexing*, Bordeaux, France, pp. 328–335.

Hopfgartner, F., Urruty, T., Villa, R., Gildea, N., and Jose, J. M. (2008a). Exploiting Log Files in Video Retrieval. In *JCDL '08: Joint Conference on Digital Libraries*. P. 454, 06 2008.

Hopfgartner, F., Hannah, D., Gildea, N., Jose, J. M. (2008b). Capturing Multiple Interests in News Video Retrieval by Incorporating the Ostensive Model. In *Proceedings of the Second International Workshop on Personalized Access, Profile Management, and*

Context Awareness in Databases, Auckland, New Zealand, pp. 48-55, 08 2008.

Huang, C.-W. (2003). Automatic Closed Caption Alignment Based on Speech Recognition Transcripts. *ADVENT Technical report*, University of Columbia.

Jaimes, A., Christel, M., Gilles, S., Ramesh, S., and Ma, W.-Y. (2005). Multimedia Information Retrieval: What is it, and why isn't anyone using it? In *MIR '05: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, pp. 3–8. ACM Press.

Joho, H., Birbeck, R. D., Jose, J. M. (2007). An Ostensive browsing and Searching on the Web. In *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*, pp. 81-92. Roskilde University Research Report.

Kelly, D. and Teevan, J. (2003). Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum* 32(2).

Kraaij, W., Westerveld, T., and Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, pp. 27–34. ACM Press.

Lee, H., Smeaton, A. F., O'Connor, N. E., and Smyth, B. (2006). User Evaluation of Físchlár - News: An Automatic Broadcast News Delivery system. *ACM Trans. Inf. Syst.* 24(2), 145–189.

Morrison, S. and Jose, J. (2004). A comparative study of online news retrieval and presentation strategies. In *ISMSE '04: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, Washington, DC, USA, pp. 403 409. IEEE Computer Society.

O'Toole, C., Smeaton, A., Murphy, N., and Marlow, S. (1999, 02). Evaluation of Automatic Shot Boundary Detection on a Large Video Test Suite. In *Proceedings of Challenges in Image Retrieval*, Newcastle, UK.

Pickering, M. J., Wong, L., and Rüger, S. (2003). ANSES: Summarisation of news video. *Image and Video Retrieval* 2788, 481–486.

Rautiainen, M., Ojala, T., and Seppänen, T. (2005). Content-based Browsing in Large News Video Databases. In *Visualization, Imaging, and Image Processing*.

Robertson, S. E., Walker, S., Jones, S., Hancock- Beaulieu, M., and Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA.

Sebe, N. and Tian, Q. (2007). Personalized Multimedia Retrieval: The New Trend? In *MIR '07: Proceedings of the International Workshop on Multimedia Information Retrieval*, New York, NY, USA, pp. 299–306. ACM Press.

Smeaton, A. F. (2002). The Físchlár Digital Library: Networked Access to a Video Archive of TV News. In *TERENA Networking Conference 2002*, Limerick, Ireland, 3-6 June 2002.

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation Campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, pp. 321–330. ACM Press.

Snoek, C. G. M., Worring, M., Koelma, D. C., and Smeulders, A. W. M. (2007). A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Transactions on Multimedia* 9(2), 280–292.

Snoek, C. G. M., Worring, M., van Gemert, J., Geusebroek, J.-M., Koelma, D., Nguyen, G. P., de Rooij, O., and Seinstra, F. (2005). MediaMill: Exploring News Video Archives based

on Learned Semantics. In *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*, New York, NY, USA, pp. 225–226. ACM Press.

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

Urban, J., Jose, J. M., van Rijsbergen, C. J. (2003). An Adaptive Approach Towards Content-Based Image Retrieval. In *Proceedings of the Third International Workshop on Content-based Multimedia Indexing*, pp. 119-126.

Vallet, D., Hopfgartner, F., and Jose, J. (2008). Use of Implicit Graph for Recommending Relevant Videos: A Simulated Evaluation. In *ECIR'08 - Proceedings of the 30th European Conference on Information Retrieval*, Glasgow, United Kingdom. Springer Verlag.

White, R., Bilenko, M., and Cucerzan, S. (2007). Studying the use of popular destinations to enhance web search interaction. In *ACM SIGIR '07: Proceedings of the 30th International ACM SIGIR Conference*, Amsterdam, The Netherlands, pp. 159–166. ACM Press.

White, R., Jose, J., van Rijsbergen, C., and Ruthven, I. (2004). A Simulated Study of Implicit Feedback Models. In *ECIR'04: Proceedings of the 26th European Conference on Information Retrieval Research*. Springer Verlag.

White, R. W., Jose, J. M., and Ruthven, I. (2003). An approach for implicitly detecting information needs. In *CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management*, New York, NY, USA, pp. 504–507. ACM.