Contextual Video Recommendation by Multimodal Relevance and User Feedback

TAO MEI, Microsoft Research Asia
BO YANG, University of Southern California
XIAN-SHENG HUA and SHIPENG LI, Microsoft Research Asia

With Internet delivery of video content surging to an unprecedented level, video recommendation, which suggests relevant videos to targeted users according to their historical and current viewings or preferences, has become one of most pervasive online video services. This article presents a novel contextual video recommendation system, called VideoReach, based on multimodal content relevance and user feedback. We consider an online video usually consists of different modalities (i.e., visual and audio track, as well as associated texts such as query, keywords, and surrounding text). Therefore, the recommended videos should be relevant to current viewing in terms of multimodal relevance. We also consider that different parts of videos are with different degrees of interest to a user, as well as different features and modalities have different contributions to the overall relevance. As a result, the recommended videos should also be relevant to current users in terms of user feedback (i.e., user click-through). We then design a unified framework for VideoReach which can seamlessly integrate both multimodal relevance and user feedback by relevance feedback and attention fusion. VideoReach represents one of the first attempts toward contextual recommendation driven by video content and user click-through, without assuming a sufficient collection of user profiles available. We conducted experiments over a large-scale real-world video data and reported the effectiveness of VideoReach.

Categories and Subject Descriptors: H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; I.2.10 [**Computing Methodologies**]: Vision and Scene Understanding—*Video analysis*

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: Video recommendation, image retrieval, relevance feedback

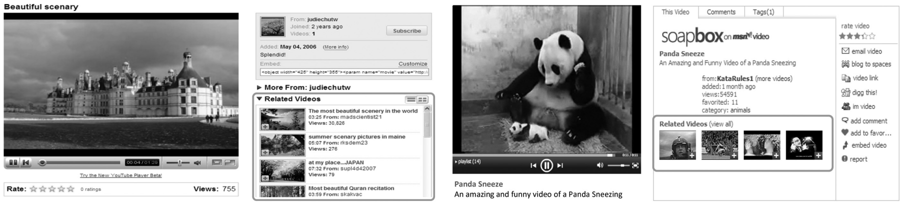
ACM Reference Format:
Mei, T., Yang, B., Hua, X-S., and Li, S. 2011. Contextual video recommendation by multimodal relevance and user feedback. ACM Trans. Inf. Syst. 29, 2, Article 10 (April 2011), 24 pages.
DOI = 10.1145/1961209.1961213 http://doi.acm.org/10.1145/1961209.1961213

1. INTRODUCTION

Driven by the age of Internet generation (especially along with the so-called Web 2.0 wave) and the advent of near-ubiquitous broadband Internet access, online delivery of video has surged to an unprecedented level. This trend has brought a wide variety of online video services, such as search, editing, sharing, advertising, and so on

Authors' addresses: T. Mei (corresponding author), X.-S. Hua, S. Li, Microsoft Research Asia, Building 2, No. 5 Dan Ling Street, Haidian District; Beijing, 100080 P. R. China; email: tmei@microsoft.com; B. Yang, Institute for Robotics and Intelligent Systems, University of Southern California, Los Angeles, CA 90089.

DOI 10.1145/1961209.1961213 http://doi.acm.org/10.1145/1961209.1961213


Beautiful scenary

From: judiechutw
Joined: 2 years ago
Videos: 1

Added: May 04, 2006 (More info)
Splendid!

Embed:
<object width="425" height="355"><param name="movie" value="http://

More From: judiechutw

Related Videos

The most beautiful scenery in the world
03:25 From: madscientist21
Views: 30,826

summer scenary pictures in maine
05:07 From: rksdem23
Views: 276

at my place...JAPAN
07:32 From: supl4d42007
Views: 79

Most beautiful Quran recitation
03:59 From: skakvac

Rate: ☆☆☆☆☆ 0 ratings Views: 755

Panda Sneeze
An amazing and funny video of a Panda Sneezing

soapbox on msn video
Panda Sneeze
An Amazing and Funny Video of a Panda Sneezing

from: KataRules1 (more videos)
added: 1 month ago
views: 54591
favorited: 11
category: animals

Related Videos (view all)

rate video
★★★☆☆

(a) Youtube (b) MSN Video

Fig. 1. The representative existing video recommender systems. The “Related Videos” in the highlighted rectangles will appear as the recommended videos after the current videos are viewed. The recommended videos are generally related to the associated text rather than the video content.

[Boll 2007]. It is natural to imagine that today’s online users always face a daunting volume of video contents, be it from video sharing or blog content, or from IPTV and mobile TV. As a result, there is an increasing demand of video recommendation services to push “interesting” or “relevant” content to targeted people at every opportunity. Video recommendation can release users’ efforts on manually filtering out unrelated content and finding the most interesting videos according to their historical and current viewings or preferences.

Many existing video-oriented sites, such as YouTube [2011], MSN Video [2011] and Yahoo! Video [Yahoo! 2011], have already provided recommendation services. Figure 1 shows some representative recommendation systems. Though no details of the algorithms in these systems are available, it is likely that most of them recommend videos only based on the relevance derived from textual information (such as query and keywords) or collaboratively based on user preference (if he/she is a registered user). As a result, the recommended videos are generally related to the associated text rather than the video content. However, user-provided tags (i.e., keywords and surrounding text) are typically incomplete and noisy. Heavy reliance on these tags will lead to the recommended videos not reaching the content of the clicked video. On the other hand, in many real cases, a user visits a Web page anonymously and is less likely to provide his/her profile. Therefore, a recommendation method without the assumption of a large collection of user profiles is highly desired. Moreover, few existing recommendation systems take users’ click-through (e.g., which part of video has been watched differently, the transition between two clicked videos, etc.) into consideration. It still remains a challenging problem to leverage video content and user click-though for a more effective recommendation.

Earlier research on recommendation began in the mid-1990’s [Adomavicius and Tuzhilin 2005], in which a recommender system is defined as estimating *ratings* for unseen items by a user. Resnick and Varian has given a more general definition as assisting and augmenting the natural social process [Resnick and Varian 1997]. There are three major types of approaches commonly used in conventional recommender systems [Adomavicius and Tuzhilin 2005]: (1) *collaborative filtering*, which predicts the preference of a given user based on the ratings of users whose tastes are similar to the given user [Fouss et al. 2007], (2) *content-based* recommendation, which predicts the interest of a user on a given item based on the similarity between the given item description and the user profile, or between the given item and items which the user has already rated [Gibas et al. 2008], and (3) *hybrid* recommendation in which both of the preceding two approaches are used [Iwata et al. 2008]. In general, collaborative filtering focuses more on the user-user relationship but less on properties of recommended items, and therefore it often requires sufficient existing users and their profiles. On the other hand, content-based recommendation focuses more on the item-user relationship but less on influence of other users. In this article, we predominantly discuss the specific domain of online video in which the input of a video recommender system is the given video clicked by a user, together with associated information such as query and surrounding texts provided by publishers, as well as how the user interacts with the videos, while the output is a list of recommended videos according to current viewing and user click-through. Specifically, we focus on the content-based technique which is more general and suitable for online video recommendation.

There exists rich research on video recommendation [Baluja et al. 2008; Bollen et al. 2005; Christakou and Stafylopatis 2005; Setten and Veenstra 2003]. Most of the these works have mainly focused on *collaborative filtering*, that is, suggesting a personalized list of videos based on collaborative ratings or user profiles, with the assumption that a sufficient collection of user ratings or profiles is available. However, as we mentioned before, users tend to visit a Web page anonymously, and therefore it is usually difficult to get user profiles. However, user click-through history of certain videos can tell underlying user interest and thus can be leveraged to find video targeted to the user [Hu et al. 2007]. For example, user's browsing behavior (e.g., the staying duration) on a specific video indicates whether this video is of interest. Furthermore, a video is a long sequence with diverse contents. When watching a video, the user may pause to have a close-up view of certain objects, fast forward to seek interesting segments, or stop browsing current video and click another new video. These user click-through data can be mined to estimate relevance and user interest.

Therefore, there are two critical issues in a video recommendation system: (1) how to fuse the relevance from multimodal information such as textual, visual, and aural modalities, and (2) how to provide personalized recommendation according to user's click-through data. Motivated by the previous observations, we build upon our previous work on video recommendation [Mei et al. 2007c; Yang et al. 2007] and propose a comprehensive online video recommendation system called VideoReach, which leverages multimodal content relevance and user click-through data for contextual recommendation. In VideoReach, a video is represented as a video document which consists of not only video (i.e., visual and aural content), but also related textual information (such as the query, keywords, and surrounding text, as well as transcript or automatically speech recognized results). Given a video document selected by a user, the recommender aims to find a list of the most contextually relevant videos based on multimodal relevance and user feedback. We believe that the relevance between two video documents should be described not only based on textual relevance, but also based on visual and aural relevance. To efficiently combine the relevance from three modalities, we adopt an Attention Fusion Function (AFF) which was successfully used in multimedia information retrieval by exploiting the variance among multimodal relevance. Furthermore, we use Relevance Feedback (RF) to automatically adjust the intraweights within each modality and the interweights among different modalities based on user click-through. We also design a voting approach by tracking the user's specific browsing behaviors on video shots to estimate the feature weights associated with the individual shots. Experiments have indicated that VideoReach outperforms one commercial video recommendation system and our previous work [Mei et al. 2007c; Yang et al. 2007].

The contributions of this article can be summarized as follows.

- We propose a systematic framework for video recommendation which is dedicated to video rather than general Web pages.
- We use multimodal content-based relevance to rate videos against the given video, as well as a voting-based approach and an attention fusion function to integrate the relevance.

- We leverage user click-through data and adopt relevance feedback and a voting approach to automatically adjust the different contributions of relevance from different video segments, features, and modalities. In this way, the system is able to track users' click-through and provide personalized recommendation.

The contextual video recommendation system proposed in this work moves one step forward from our previous effort [Mei et al. 2007c; Yang et al. 2007] in that: (1) the recommendation problem is formulated as finding a set of suitable functions $\{g_i\}$ and feature weights $\{w_i, w_{ij}, w_{ijk}\}$ for computing multimodal relevance; (2) it leverages users' browsing behaviors on specific video segments to estimate the corresponding feature weight, which has significantly improved recommendation performance; (3) visual concept detection is used to estimate the video category, which also improves the recommendation performance; and (4) the system is evaluated through more comprehensive experiments.

The rest of this article is organized as follows: Section 2 reviews related work on video recommendation; the system framework of VideoReach is given in Section 3. We detail textual, visual, and aural relevance in Section 4; fusion strategies of multimodal relevance and relevance feedback are presented in Section 5. Section 6 presents experimental results, followed by conclusions in Section 7.

2. RELATED WORK

The research problems closely related to video recommendation are conventional recommendation, video search, and video content understanding. We will briefly review previous work on these topics.

2.1. Conventional Recommendation

As we have mentioned in Section 1, research on conventional recommendation has proceeded along three major dimensions, that is, *content-based*, *collaborative*, and *hybrid* recommendation. Additionally, Burke has summarized the techniques in traditional recommenders using another three dimensions as follows [Burke 2002].

- *Demographic-based* recommendation: to collect sufficient users' information from diverse sources such as interactive responses and users' homepages, create relationships among users based on this information, and finally make recommendations like *collaborative* recommendation.
- *Utility-based* recommendation: to let users input a "utility function" to directly describe their needs.
- *Knowledge-based* recommendation: to aggregate the knowledge not only from average users but also from experts in a specific area for recommendation. An example of this approach is the Encyclopedia system [2011].

To summarize, the first dimension proposed by Burke (i.e., demographic-based recommendation) heavily relies on a sufficient collection of user profiles. Although this information is not required in the latter two, it is hard to choose a general "utility function" satisfying all users. Moreover, when surfing the Web, most users browse Web pages anonymously. As a result, it is a common sense that an online video recommender should deal with the absence of user profiles. Therefore, conventional recommender techniques cannot be directly applied in online video. A more comprehensive literature survey on recommender systems can be found in Adomavicius and Tuzhilin [2005].

Many recommendation systems have been developed in a variety of areas, such as text documents [Zhou et al. 2008], movies [Christakou and Stafylopatis 2005], TVs [Setten and Veenstra 2003], Web pages [Balabanovic 1998], digital libraries [Bollen

Table I. Comparisons between Video Search and Video Recommendation

	Keyword query	Visual-aural query	Relevance	Collaborative rating	Click-thru
Video Search	√	×	√	×	×
Video Recommendation	√	√	√	√	√

"√" denotes "used" while "×" denotes "not used"

et al. 2005], online marketing [Wei et al. 2005], online videos [Baluja et al. 2008], and so on. It is observed that most of these recommenders assume that a sufficient collection of user profiles is available. In other words, these approaches mainly focus on *collaborative* recommendation based on user profiles. In general, user profiles mainly come from two types of sources: (1) direct profile, that is, user's selection of a list of predefined interests; (2) indirect profile, that is, user's rating of a number of items. Regardless of what kinds of items are recommended by these systems, the objective is to collaboratively recommend the items matching to the profile or interest. However, *content-based* recommendation dedicated to video has not yet been deeply studied.

2.2. Video Search

The techniques used in video search can be classified into three categories, that is, text-based, content-based, and multimodal-based approaches [Chang et al. 2007]. Text-based video search aims to find relevant videos according to textual relevance, content-based search leverages visual content similarity for searching visually or conceptually similar videos [Hauptmann et al. 2008], while the multimodal-based approach uses the first two in a hybrid way. Recently, a great deal of effort has been carried out on multimodal-based video search [Kennedy et al. 2008; Mei et al. 2007a], where content features can be used directly to compute the similarity between videos, or used with users' interactive evaluations, or used for reranking the results returned by text-based search.

It is worth noting that an alternative to video recommendation is to adopt the techniques used in the domain of video search. However, the tasks of video search and video recommendation are different. Video search aims to find videos that exactly "match" the given query. In other words, the "relevance" in video search usually indicates exact "match," while it indicates not only "match" but also with common "interest." Such interest can be derived from collaborative ratings from a set of uses with similar profiles or rating the same videos, or from user click-through. In addition, the query in a search system is typically a list of keywords or is given together with an example of an image (although query by example is not yet realized by any search engine so far), while the query in a recommender system is a video document clicked or searched by a user with certain interest and click-though. This query may consist of visual-aural content and textual information associated with this video. Therefore, multimodal content relevance, as well as user feedback, should be taken into account for video recommendation. Table I lists the comparisons between video search and recommendation systems. It can be concluded that video recommendation is more general than video search.

2.3. Video Content Understanding

Another research topic related to video recommendation is video content understanding, which aims to understand visual-aural content in the video stream. The key idea is to map the video content to textual descriptions (e.g., a set of keywords). Once we can understand video content (i.e., describe video content using textual

Table II. Key Notations in the VideoReach System

U	user
T, V, A	textual, visual, and aural modalities in a video
i	index of modality, $i \in \{T, V, A\}$
j	index of feature
k	index of shot within a video
d_i	i-th modality in a video
D	video document, $D \triangleq (d_V, d_T, d_A)$
D_x, D_y	video document D_x and D_y
$\mathbf{f}_i$	feature vector for i-th modality, $\mathbf{f}_i = (f_{i1}, \ldots, f_{ij}, \ldots)$
f_{ij}	j-th feature from i-th modality in a video
f_{ijk}	j-th feature from i-th modality in k-th shot
$R(D_x, D_y)$	relevance between video D_x and D_y, which is used for recommendation
$R_i(D_x, D_y)$	relevance between video D_x and D_y in terms of i-th modality
$R_{ij}(D_x, D_y)$	relevance between video D_x and D_y in terms of j-th feature from i-th modality
ω_i	weight of the relevance $R_i(D_x, D_y)$
ω_{ij}	weight of the relevance $R_{ij}(D_x, D_y)$
ω_{ijk}	weight of feature f_{ijk} (the degree of interest of k-th shot for current user), note that ω_{ijk} is dependent to a document D_x
g_1, g_2, g_3	a set of functions for computing different kinds of relevance

words), we can leverage recommendation techniques in the text domain to improve recommendation performance. This is particularly useful for content-based recommendation, since content understanding can lead to more precise and richer description of video content, which in turn benefits the computation of content-based relevance. According to the taxonomy used for describing video content, the approaches to video content understanding can be roughly categorized into two classes: (1) those heavily relying on a predefined taxonomy (i.e., a set of predefined concepts or event lexicon, such as LSCOM ontology [Naphade et al. 2006] and MediaMill 101 concepts [Snoek et al. 2006]) [Gu et al. 2008; Shen et al. 2008], and (2) the others which are independent to any taxonomy (also called "tagging") [Moxley et al. 2010; Siersdorfer et al. 2009]. More recent comprehensive surveys can be found in Datta et al. [2008], Lew et al. [2006], and Snoek and Worring [2009].

3. SYSTEM OVERVIEW

3.1. Notations and Problem Formulation

For the sake of mathematical tractability, we list the key notations in Table II. Given a clicked video document D which can be represented by the triplet of textual, visual, and aural modalities as $D \triangleq (d_T, d_V, d_A)$, the task of video recommendation is expressed as finding a list of the most relevant videos to D and the given user U, where T, V, and A denote the textual, visual, and aural modalities, respectively. The document of each individual modality d_i can be represented by a set of features $d_i \triangleq d_i(\mathbf{f}_i)$, where $\mathbf{f}_i = (f_{i1}, \ldots, f_{ij}, \ldots)$ is a vector of the features from modality i, f_{ij} denotes the jth feature from modality i. Furthermore, as different parts of video may have different degrees of interest to a user, we use shot as the basic unit of video segment and f_{ijk} to denote the jth feature from the ith modality in the kth shot, where $k = \{1, \ldots, |D|\}$ and

$|D|$ is the number of shots in the video D.[1] Accordingly, we can use ω_{ijk} $(0 \leqslant \omega_{ijk} \leqslant 1)$ to denote the weight for feature f_{ijk} in the kth shot. Then, f_{ij} could be the linear combination of f_{ijk}, given by

$$f_{ij} = \sum_{k=1}^{|D|} \omega_{ijk} \cdot f_{ijk}. \tag{1}$$

The preceding notations can be explained through the following examples. The features of the textual modality $\mathbf{f}_T$ could consist of the *term frequency* (*tf*) and *inverted document frequency* (*idf*) which are widely adopted in information retrieval [Baeza-Yates and Ribeiro-Neto 1999]. The features of visual modality $\mathbf{f}_V$ could be represented by *color*, *texture*, *motion*, and so on, each corresponding to a set of elements in the feature vector $\mathbf{f}_V$. The value f_{ij} of a motion feature is the jth element in $\mathbf{f}_V$, which could be the weighted combination of all the features f_{ijk} across different shots.

Based on these notations, the problem of video recommendation can be formulated as follows. Let $R(D_x, D_y)$ denote the multimodal relevance of two video documents D_x and D_y, which is used to recommend videos, and $\mathcal{D}$ denote the video database. Given a video document D_x clicked by current user U, video recommendation is to seek the video D_y^* which satisfies

$$D_y^* = \arg \max_{D_y \in \{\mathcal{D} \backslash D_x\}} R(D_x, D_y). \tag{2}$$

Since different modalities have different contributions to the overall relevance, we can use $R_i(D_x, D_y)$ $(i \in \{T, V, A\})$ to denote the relevance between D_x and D_y in terms of the ith modality, and ω_i to denote the weight of the relevance from the ith modality. Then, the multimodal relevance $R(D_x, D_y)$ between video D_x and D_y is given by

$$R(D_x, D_y) = g_1\big(R_i(D_x, D_y), \omega_i\big), \tag{3}$$

where g_1 is a function operated on R_i and ω_i, and will be explained later.

To obtain $R_i(D_x, D_y)$, we need to consider the relevance from different types of feature in modality i. Let $R_{ij}(D_x, D_y)$ denote the relevance in terms of the jth feature within the ith modality. Similarly, $R_i(D_x, D_y)$ can be obtained by a function g_2 on R_{ij} and the corresponding weight ω_{ij} by

$$\begin{aligned} R_i(D_x, D_y) &= g_2\big(R_{ij}(D_x, D_y), \omega_{ij}\big) \\ &= \sum_j \omega_{ij} R_{ij}(D_x, D_y). \end{aligned} \tag{4}$$

Given the feature $f_{ij}(D_x)$ and $f_{ij}(D_y)$, the relevance $R_{ij}(D_x, D_y)$ between documents D_x and D_y in terms of the jth feature in the ith modality can be obtained by a distance function g_3 as

$$\begin{aligned} R_{ij}(D_x, D_y) &= g_3\big(f_{ij}(D_x), f_{ij}(D_y)\big) \\ &= g_3\Big(\sum_k \omega_{ijk} f_{ijk}(D_x), \sum_k \omega_{ijk} f_{ijk}(D_y)\Big), \end{aligned} \tag{5}$$

[1] A shot is defined as an uninterrupted temporal segment in a video, recorded by a single camera.

where g_3 measures the distance or the similarity between two feature sets $f_{ij}(D_x)$ and $f_{ij}(D_y)$ of the document D_x and D_y. Therefore, Eq. (3) can be rewritten as the operations of the three functions g_1, g_2, and g_3 by

$$\begin{aligned} R(D_x, D_y) &= g_1\big(R_i(D_x, D_y), \omega_i\big) \qquad (6)\\ &= g_1\Big(g_2\big(R_{ij}(D_x, D_y), \omega_{ij}\big), \omega_i\Big)\\ &= g_1\Bigg(g_2\Big(g_3\big(\sum_k \omega_{ijk}\, f_{ijk}(D_x), \sum_k \omega_{ijk}\, f_{ijk}(D_y)\big), w_{ij}\Big), w_i\Bigg). \end{aligned}$$

Eqs. (2) and (6) give the formulation of video recommendation in VideoReach. Given the feature f_{ijk} (i—modality, j—feature, k—shot) of two videos, we aim to obtain a set of feature weights $(\omega_i, \omega_{ij}, \omega_{ijk})$ and functions (g_1, g_2, g_3), so that we can use $R(D_x, D_y)$ to rank the recommended videos. The strategies for selecting different functions of (g_1, g_2, g_3) are different. Regarding g_3, since each modality has its unique content characteristics, we adopt different functions for computing the relevance within textual, visual, and aural modalities. Regarding g_2, since linear combination is a straightforward yet most effective approach in multimedia information retrieval, we used linear weighted fusion for combining the relevance from different modalities. To deal with the problem of the conformance with human perception for different modality in the linear fusion, we adopt the Attention Fusion Fuction (AFF) which has been shown effective as g_1 [Hua and Zhang 2004; Hua et al. 2004a]. In the proposed VideoReach system, we adopted different distance metrics (e.g., L_1 distance and vector space model) for g_3 based on different types of features, a linear weighted fusion function for g_2, and an attention fusion function for g_1. Then, the key problem is how to compute the feature weights $(\omega_i, \omega_{ij}, \omega_{ijk})$ in Eq. (6). The decision of these weights is based on user feedback, that is, user click-through of certain videos and video segments. Given a user, as different segments in a video have different degrees of interest, and as different modalities have different contributions to the overall relevance, we adopt a voting-based approach to compute ω_{ijk} and relevance feedback technique to compute ω_i and ω_{ij} [Rui et al. 1998].

3.2. System Framework

Figure 2 illustrates the system framework of the proposed VideoReach system. First, based on user click-through of a given video (i.e., user's browsing behavior on this video), each shot is assigned with a feature weight ω_{ijk}, reflecting the degree of interest of this user to this segment. The feature f_{ij} is obtained by the weighted combination of feature f_{ijk} in the kth shot (refer to Eq. (1)). The relevance R_{ij} for a specific feature in a single modality is obtained by the function g_3 operated on the feature f_{ij} (refer to Eq. (5)). As we have mentioned, g_3 could be based on specific distances (e.g., L_1 distance or vector-based distance). Similarly, the relevance R_i in terms of a single modality i is computed by weighted linear combination of relevances from all the features within this modality (refer to Eq. (4)). Then, the relevances from all modalities are fused using the AFF [Hua and Zhang 2004; Hua et al. 2004a]. The intraweights ω_{ij} within each modality and interweights ω_i among different modalities are dynamically adjusted using the Relevance Feedback (RF) technique [Rui et al. 1998] based on user click-through history (i.e., whether the user paid enough attention to this video and what is the next video he/she clicks). The feature weight ω_{ijk} in the kth shot is updated according to user's browsing behavior (whether this user played, skipped, or fast browsed this shot). VideoReach runs in an iterative way so that the relevance of the recommended videos can be improved to be less intrusive and more targeted to the given user.

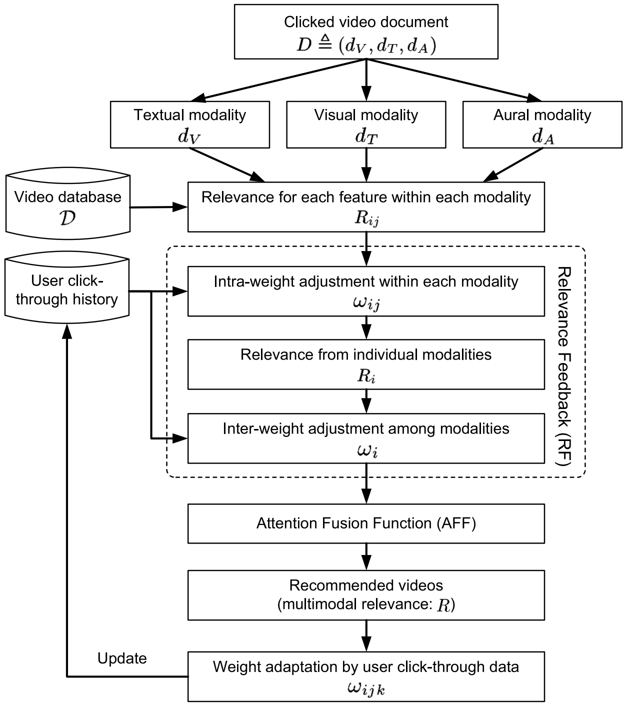

Clicked video document
$D \triangleq (d_V, d_T, d_A)$
Textual modality
d_V
Visual modality
d_T
Aural modality
d_A
Video database
$\mathcal{D}$
Relevance for each feature within each modality
R_{ij}
User click-through history
Intra-weight adjustment within each modality
ω_{ij}
Relevance from individual modalities
R_i
Inter-weight adjustment among modalities
ω_i
Relevance Feedback (RF)
Attention Fusion Function (AFF)
Recommended videos
(multimodal relevance: R)
Update
Weight adaptation by user click-through data
ω_{ijk}

Fig. 2. System framework of VideoReach.

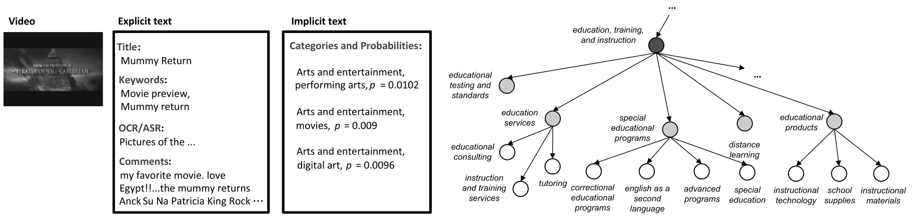
It is worth noting that although using textual features to compute the relevance of video documents is the most common method and can work well in most cases, not all concepts can be well described by text only. For example, for a video about "beach," the keywords related to "beach" may be "sky," "sand," "people," and so on. Meanwhile, these words are probably related to many other unrelated videos, such as "desert," "weather," and so on. It is reasonable to use visual features to describe "beach" rather than text. Furthermore, aural features are also important for the overall relevance, especially for music videos. Therefore, in addition to textual features, we use visual and aural features to augment the textual description of video. We next describe the relevance from textual, visual, and aural modalities, as well as fusion strategy by AFF and RF.

4. MULTIMODAL RELEVANCE

Video is a compound of image sequence, audio track, and textual information, which each deliver information with their own primary elements. Accordingly, multimodal content relevance is represented by the combination of relevances from the three modalities. We will detail textual, visual, and aural relevances in this section.

4.1. Textual Relevance

We classify textual information related to a video document into two categories: (1) *explicit* text, referring to the surrounding text provided by publisher, transcript, Automated Speech Recognition (ASR) results, and Optical Character Recognition (OCR) embedded in the video stream; (2) *implicit* text, referring to the (hidden) categories and their probabilities obtained by automatic text categorization based on a set of predefined taxonomies (i.e., category hierarchy). Figure 3(a) shows an example of the



Title:
Mummy Return
Keywords:
Movie preview,
Mummy return
OCR/ASR:
Pictures of the ...
Comments:
my favorite movie. love
Egypt!!...the mummy returns
Anck Su Na Patricia King Rock ···

Categories and Probabilities:
Arts and entertainment, performing arts, p = 0.0102
Arts and entertainment, movies, p = 0.009
Arts and entertainment, digital art, p = 0.0096

...
education, training, and instruction
...
educational testing and standards
education services
special educational programs
distance learning
educational products
educational consulting
instruction and training services
tutoring
correctional educational programs
english as a second language
advanced programs
special education
instructional technology
school supplies
instructional materials

(a) an example of explicit and implicit texts (b) a part of taxonomy

Fig. 3. An example of the textual information associated with a video.

explicit and implicit textual information associated with a video. We use the Vector Space Model (VSM) and probabilistic model to describe the explicit and implicit texts, respectively [Yang et al. 2007]. As a result, the textual modality d_T is represented as the combination of two kinds of features $d_T \triangleq (\mathbf{f}_T^{(E)}, \mathbf{f}_T^{(I)})$, where $\mathbf{f}_T^{(E)}$ and $\mathbf{f}_T^{(I)}$ denote the feature vector for the explicit and implicit texts, respectively.

In the VSM, each element of $\mathbf{f}_T^{(E)}$ is the weight for the corresponding word appearing in the whole video database $\mathcal{D}$. Note that for the sake of simplicity, we drop $\mathcal{D}$ from the notations $\mathbf{f}_T^{(E)}$. Here, we can use the term frequency tf to describe the weight of a word and *cosine* similarity as the measurement of relevance between two videos D_x and D_y in terms of explicit text [Baeza-Yates and Ribeiro-Neto 1999].

$$R_{T1}(D_x, D_y) = \frac{\mathbf{f}_T^{(E)}(D_x) \cdot \mathbf{f}_T^{(E)}(D_y)}{||\mathbf{f}_T^{(E)}(D_x)|| \times ||\mathbf{f}_T^{(E)}(D_y)||} \tag{7}$$

Note that $R_{T1}(D_x, D_y)$ indicates the relevance in terms of the first feature from the textual modality (corresponding to $i = T, j = 1$ in R_{ij}, refer to the definition of R_{ij} in Table II).

Although the vector model is able to present the keywords of a textual document, it is not enough to describe latent semantics in the videos. For example, for an introduction to a music video named “flower,” “flower” is an important keyword and has a high weight in the VSM. Consequently, many videos related to real flowers will be recommended by Eq. (7). However, videos related to music are more relevant. To address this problem, we adopt a probabilistic model as a complement to the VSM [Yang et al. 2007]. Specifically, we build a taxonomy of keywords and leverage the category information in this taxonomy obtained by a probabilistic model. This taxonomy is first built based on the 67 target categories (from KDDCUP 2005[2]), and then enriched based on the queries from a commercial search engine [Shen et al. 2006a, 2006b]. Figure 3(b) shows a part of the taxonomy. We adopt text categorization techniques based on Support Vector Machines (SVMs) [Yang and Liu 1999] to classify a textual document into a set of this hierarchical taxonomy. The relevance of two documents in this probabilistic model $R_{T2}(D_x, D_y)$ $(i = T, j = 2)$ is the similarity between the corresponding hidden category hierarchy $\mathbf{f}_T^I$ [Yang et al. 2007]. As a result, the textual relevance can be computed by a linear combination of the relevances of $R_{T1}(D_x, D_y)$ and $R_{T2}(D_x, D_y)$. We will describe later how we can get the corresponding weights for relevance fusion.

[2] http://www.acm.org/sigkdd/kddcup

4.2. Visual Relevance

The visual relevance is measured by *color*, *motion*, and *shot tempo* (the average number of shots per second), which had proved effective to describe dominant visual content in many existing video retrieval systems [Hua et al. 2004a; Mei et al. 2007a]. Specifically, we use the normalized 64-dimensional color histogram in the HSV space and motion intensity which is computed by the frame-difference to represent *color* and *motion* features, respectively. For more details about these visual features, please refer to Mei et al. [2007a]. Furthermore, we use the automatically recognized video *concepts* as a type of high-level visual feature, as they have proven effective for video retrieval [Hauptmann et al. 2008]. We used the 36 concepts defined in TRECVID 2007 corpus [TRECVID 2011].[3] As a result, the visual modality d_V can be represented as $d_V \triangleq d_V(\mathbf{f}_{V1}, \mathbf{f}_{V2}, \mathbf{f}_{V3}, \mathbf{f}_{V4})$, where $\mathbf{f}_{V1}$, $\mathbf{f}_{V2}$, $\mathbf{f}_{V3}$, and $\mathbf{f}_{V4}$ represent the feature vectors of *color* (64-dimensional), *motion* (1-dimensional), *shot tempo* (1-dimensional), and *concepts* (36-dimensional), respectively. For two video documents D_x and D_y, the visual relevance $R_{Vj}(D_x, D_y)$ $(i = V)$ in terms of the *j*th feature $(j = 1, 2, 3)$ is defined based on L_1 distance, that is, g_3 is a function operated on L_1 distance as follows. We have

$$\begin{aligned} R_{Vj}(D_x, D_y) &= g_3\big(\mathbf{f}_{Vj}(D_x), \mathbf{f}_{Vj}(D_y)\big) \\ &= 1.0 - L_1\big(\mathbf{f}_{Vj}(D_x), \mathbf{f}_{Vj}(D_y)\big) \\ &= 1.0 - \big|\mathbf{f}_{Vj}(D_x) - \mathbf{f}_{Vj}(D_y)\big|, \end{aligned} \tag{8}$$

where the modality index $i = V$.

Intuitively, the more similar the two depicted videos in terms of appearance, the more relevant they are. However, as we have mentioned, a video document can also be represented by the presence of a set of predefined high-level visual concepts $\mathbf{f}_{V4}$. Then, the similarity between two videos can be computed by the intersection of the corresponding concept vectors [Mei et al. 2007b]. The visual relevance in the concept space between two video documents D_x and D_y can be given by

$$\begin{aligned} R_{V4}(D_x, D_y) &= g_3\big(\mathbf{f}_{V4}(D_x), \mathbf{f}_{V4}(D_y)\big) \\ &= 1.0 - \frac{\sum_{t=1}^{T} \min\big(f_{V4}^{(t)}(D_x), f_{V4}^{(t)}(D_y)\big)}{\sum_{t=1}^{T} f_{V4}^{(t)}(D_y)}, \end{aligned} \tag{9}$$

where g_3 is defined as the “intersection” operation here, $f_{V4}^{(t)}$ indicates the *t*th element of vector $\mathbf{f}_{V4}$, and $T = 36$ is the number of concepts.

4.3. Aural Relevance

The aural modality d_A is described using the linear weighted fusion of aural tempos among all the shots. In addition, we also use the standard deviation of the tempos from all the shots. The average aural tempo represents the speed of music or audio, while the standard deviation indicates the change frequency of music style. These features have proved effective to describe aural content [Hua et al. 2004a; Shen et al. 2009]. Specifically, the audio track is segmented into audio clips by detecting strong beats. A strong beat is taken as the boundary of an audio clip. Then, the tempo of each audio clip is estimated by the onset frequency in the clip. The higher the value, the faster the tempo. Then, the clip boundaries are aligned with shot boundaries. Usually,

[3] These concepts include: Sports, Weather, Court, Office, Meeting, Studio, Outdoors, Building, Desert, Vegetation, Mountain, Road, Sky, Snow, Urban, Waterscape_Waterfront, Crowd, Face, Person, Police_Security, Military, Prisoner, Animal, Computer_TV-screen, Flag-US, Airplane, Car, Bus, Truck, Boat_Ship, Walking_Running, People-Marching, Explosion_Fire, Natural-Disaster, Maps, Charts.

Table III. Update of $\omega_{ijk}^{(t)}$ ($\omega_{ijk}^{(0)} \equiv 1.0$)

Update	User behavior
$\omega_{ijk}^{(t+1)} = \omega_{ijk}^{(t)}$	normal browse
$\omega_{ijk}^{(t+1)} = \omega_{ijk}^{(t)} + 0.5$	pause and then browse
$\omega_{ijk}^{(t+1)} = \omega_{ijk}^{(t)} + 1.0$	seek or replay
$\omega_{ijk}^{(t+1)} = \omega_{ijk}^{(t)} - 0.5$	fast browse or skip

a shot contains multiple audio clips. For more details of audio tempo, please refer to Hua et al. [2004a].

As a result, an aural document d_A is represented by $d_A \triangleq d_A(f_{A1}, f_{A2})$, where f_{A1} and f_{A2} represent the average and the standard deviation of aural tempo, respectively. Similar to the visual relevance described in Section 4.2, the aural relevance $R_{Aj}(D_x, D_y)$ in terms of the jth feature is given based on L_1 distance, that is, g_3 is defined as L_1 distance. We have

$$\begin{aligned} R_{Aj}(D_x, D_y) &= g_3\big(f_{Aj}(D_x), f_{Aj}(D_y)\big) \\ &= 1.0 - \big|f_{Aj}(D_x) - f_{Aj}(D_y)\big|, \end{aligned} \quad (10)$$

where $j = 1, 2$. Intuitively, the more similar the two videos sound in terms of audio, the more relevant they are. The computation of the weights for fusion will be described in the next section.

5. USER FEEDBACK FOR RELEVANCE FUSION

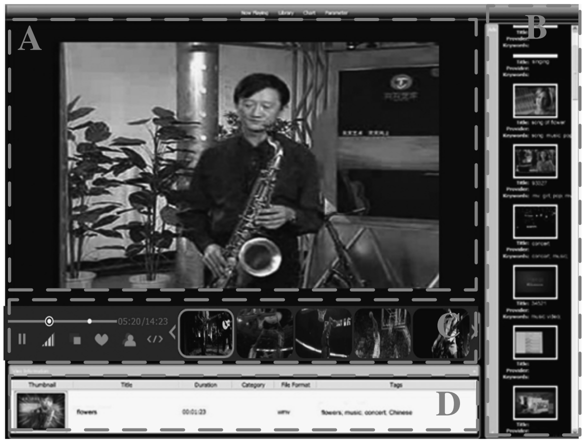
We have modeled the relevance from each individual segment (i.e., shot). However, fusing these relevances to the overall multimodal relevance for recommendation is another key issue. We will show how to combine the relevance from individual segments, feature types, and modalities by the voting-based approach, relevance feedback, and attention fusion function in this section. Please note that all the weights are adjusted in an iterative manner.

5.1. Weights Adjustment for Shot Feature (ω_{ijk})

In our previous work [Yang et al. 2007], we simply average the features from all the shots as the feature for each single modality. However, as video is a time-evolving sequence with diverse contents, users may have different degrees of interest on different parts of video [Yu et al. 2003]. Therefore, we leverage user click-through (user browsing behaviors on a video sequence) to obtain the shot feature weight ω_{ijk}. For example, if a user fastforwards or fastbackwards a segment (i.e., shot),[4] he/she may not be interested in this shot, then the weight ω_{ijk} for shot k should be decreased; if a user seeks a specific shot or to replay a shot, he/she may have strong interest on the content of this shot, then the weight ω_{ijk} should be increased.

Based on the previous observations, similar to the user browsing logs analysis in Yu et al. [2003], we record the user browsing behaviors and classify the behaviors into four categories. The shot feature weight ω_{ijk} is then dynamically adjusted by a voting-based approach, listed in Table III. The weight $\omega_{ijk}^{(t+1)}$ for shot k at the $(t+1)$ iteration depends on that in the previous step $\omega_{ijk}^{(t)}$. The user interface for shot-based video browsing is shown in Figure 4. “Panel C” provides the functionality to navigate the video via a shot

[4] We adopt shot as the basic segment in this article, as shot is a physical segment resulting from a continuous camera operation.



Thumbnail	Title	Duration	Category	File Format	Tags
	flowers	00:01:23		wmv	flowers; music; concert; Chinese

Fig. 4. User interface of our video recommendation system. A—online video; B—recommended video list; C—shot list, the highlighted shot indicates current playing content; D—related textual description of this online video.

list. Consequently, the feature of the jth feature in the ith modality of video document D_x, that is, $f_{ij}(D_x)$, is computed by

$$f_{ij}(D_x) = \sum_{k=1}^{|D_x|} \omega_{ijk}\, f_{ijk}(D_x), \tag{11}$$

where $|D_x|$ is the number of shots in the video D_x. Please note that $\omega_{ijk}^{(t)}$ is normalized to [0, 1] by

$$\omega_{ijk}^{(t)} = \frac{\omega_{ijk}^{(t)} - \omega_{\min}^{(t)}}{\omega_{\max}^{(t)} - \omega_{\min}^{(t)}}, \tag{12}$$

where $\omega_{\min}^{(t)}$ and $\omega_{\max}^{(t)}$ denote the minimum and maximum of $\omega_{ijk}^{(t)}$ in the tth iteration.

5.2. Weights Adjustment with Relevance Feedback (ω_{ij}, ω_i)

Before fusing the relevances from the three modalities, two issues need to be addressed: (1) how to obtain the intraweights of the relevances for each kind of feature within a single modality (i.e., ω_{T1} and ω_{T2} in textual modality, ω_{V1}, ω_{V2}, ω_{V3}, and ω_{V4} in visual modality, and ω_{A1} and ω_{A2} in aural modality); (2) how to decide the interweight (i.e., ω_T, ω_V, and ω_A) of the relevances for each modality.

In fact, it is not easy to select a set of weights satisfying all video documents. As we have discussed in Section 3, for the concept "beach," visual relevance is more important than the other two, while for the concept "Microsoft," textual relevance is more salient. Therefore, it is reasonable to assign different video documents with different intra- and interweights. It is observed that user click-through data usually tells a latent instruction to the assignment of the weights, or at least a kind of latent comment on the recommendation. If a user opens a recommended video and closes it within a short time (i.e., less than 15 seconds), probably this video is a false recommendation. We call such videos "negative" examples. On the other hand, if a user views a recommended video for a relatively long time, this video is probably a true recommendation, since this user is interested in this recommendation. We call such videos "positive" examples. Based on "positive" and "negative" examples, relevance feedback [Rui et al.

1998; Tao et al. 2006] is an effective solution to automatically adjusting the weights of different inputs, that is, intra- and interweights.

The adjustment of intraweights is to obtain the optimal weight of each kind of feature within an individual modality. Among the returned list, only the “positive” examples indicated by a user are selected to update intraweights as

$$\omega_{ij} = \frac{1}{\sigma_{ij}}, \tag{13}$$

where $i \in \{T, V, A\}$, σ_{ij} is the standard deviation of the jth feature f_{ij} in the ith modality, whose corresponding document D_x is a “positive” example. Intuitively, if all the recommended “positive” videos have similar values for the feature f_{ij}, it means that f_{ij} is a good indicator of the user’s information need. On the other hand, if the values for the feature f_{ij} are very different among the recommended “positive” videos, then f_{ij} is not a good indicator. The intraweights are then normalized between 0 and 1.

The adjustment of interweights is to obtain the optimal weight of each modality. For each modality, a recommended list $(D_{x_1}, D_{x_2}, \ldots, D_{x_N})$ is created based on the individual relevance from this modality, where N is the number of recommended videos. We first initialize $\omega_i = 0$, and then update w_i as

$$\omega_i = \begin{cases} \omega_i + 1, & \text{if } D_{x_n} \text{ is a “positive” example} \\ \omega_i - 1, & \text{if } D_{x_n} \text{ is a “negative” example} \end{cases},$$

where $i \in \{T, V, A\}$ and $n = 1, \ldots, N$. Intuitively, the more relevant the videos returned by modality i, the more salient the corresponding weight ω_i. The interweights are then normalized between 0 and 1. It is worth noting that other relevance feedback strategies such as Tao et al. [2006] can be also integrated into this framework.

5.3. Fusion with Attention Fusion Function

Given the weight and the relevance score for each individual modality, we need to combine them to produce a “final relevance” score; the higher the relevance, the higher the possibility that two documents are relevant. Linear combination of the relevance of individual modality is a straightforward and effective method for relevance fusion, based on the interweights obtained by relevance feedback in Section 5.2. However, this approach is not consistent with human attention response. To address this problem, Hua and Zhang have proposed an Attention Fusion Function (AFF) to simulate human attention characteristics [Hua and Zhang 2004]. The AFF-based fusion is applicable when two properties are satisfied: *monotonicity* and *heterogeneity*. Specifically, the first property indicates that the final relevance increases whenever any individual relevance increases, while the second indicates that if two video documents are relevant in terms of one individual modality but irrelevant in terms of the others, they are still perceived very relevant. We will show in the experiments that AFF is much more consistent with human perception if the aforesaid two properties are satisfied. Note that in the linear combination, both of the two properties cannot be satisfied.

In VideoReach, the first property is easy to be satisfied since each component does contribute to the overall relevance. For the second, however, since two video documents are not necessarily relevant even they are very similar in terms of one feature, we first fuse the preceding relevance into three channels, that is, textual, visual, and aural relevance. As textual information (usually provided by the content owners) is more precise and contains much more available semantic than visual and aural information, textual relevance is more reliable than that from visual and aural. Thus, only high relevance in the visual or aural channel but low relevance in the textual channel may not indicate two documents being relevant. This means the *heterogeneity* property will not be satisfied in such a case. For example, if two videos are showing a green plant and a green building, respectively, they may have high visual relevance. However, they may be not relevant at all since they are talking about different topics. On the other hand, the high textual relevance between two documents would indicate they are relevant. Therefore, we first filter out most documents with low textual relevance to ensure all the remaining documents are much more relevant to the given document, and then calculate the visual and aural relevance within these documents. As a result, the textual relevance will not dominate the overall relevance, and the *heterogeneity* property is also satisfied. According to AFF, if a document has high visual or aural relevance with the given video, a user will pay more attention to it than to others with moderate relevance scores. In our experiment, we filter out the documents which are not in the top 20 documents before performing AFF-based fusion.

In this way, the monotonicity and heterogeneity are both satisfied. We can use AFF to get better fusion results. Since different features should have different weights, we adopt the three-dimensional weighted AFF in Hua and Zhang [2004] to obtain the final relevance. For two documents D_x and D_y, the overall multimodal relevance is given by

$$R(D_x, D_y) = \frac{R_{avg} + \frac{1}{2(m-1)+m\gamma}\sum_i \left|m\omega_i R_i(D_x, D_y) - R_{avg}\right|}{W},$$

where

$$\begin{aligned} R_{avg} &= \sum\nolimits_i \omega_i R_i(D_x, D_y) \\ W &= 1 + \frac{1}{2(m-1)+m\gamma}\sum_i \left|1 - m\omega_i\right| \end{aligned} \tag{14}$$

and $i \in \{T, V, A\}$. m is the number of modalities ($m = 3$), ω_i is the weight of individual modality which can be obtained in Section 5.2, and γ ($\gamma > 0$) is a predefined constant which controls the effectiveness of one modality in the overall relevance. $R_i(D_x, D_y)$ is obtained by the linear fusion of R_{ij} in Eq. (4). In our implementation, γ is empirically fixed to 0.2. For more details of AFF function, please refer to Hua and Zhang [2004].

6. EXPERIMENTS

6.1. Data and Methodologies

We have collected more than 13, 000 online videos from MSN Video [2011] in the experiments. It is not reasonable to evaluate the performance of VideoReach over all these videos. Instead, we used 75 representative videos as the source videos (i.e., the seed videos which are supposed to be clicked by users and used to recommend videos). We used 15 representative textual queries to search videos from our database with 13, 000 videos. From the search results for each query, only the top five videos are selected as the source videos for evaluation. The content of the selected videos covers a diversity of genres, such as music, sports, cartoons, movie previews, persons, travel, business, food, and so on. The 15 queries consist of ten popular queries and five rare queries from a commercial video site. The selected ten popular queries include “flowers,” “cat,” “baby,” “sun,” “soccer,” “fire,” “beach,” “food,” “car,” and “Microsoft,” while the five rare queries include “cancun 2007,” “waterspouts,” “bubble,” “Greek,” and “shreck3”.[5] Figure 4 shows the user interface of VideoReach. In order to compare the performance

[5] These five queries are selected based on the statistics that the effective query number issued by users is less than 30 in a single month (May–June, 2007).

Table IV. Predefined Weights in Schemes (2)–(6)

	ω_T		ω_V				ω_A	
weight	ω_{T1}	ω_{T2}	ω_{V1}	ω_{V2}	ω_{V3}	ω_{V4}	ω_{A1}	ω_{A2}
intra	0.50	0.50	0.50	0.20	0.20	0.10	0.70	0.30
inter	0.70		0.15				0.15	

of VideoReach with MSN Video, as well as compare the effectiveness of different fusion strategies, for each source video, we recommended eight different video lists with each containing 10 videos. Theoretically, there are $6,000$ ($75 \times 8 \times 10$) videos in total for evaluation. However, as different recommendation schemes may return identical videos, there are quite a few duplicates. As a result, there are $3,512$ unique videos recommended for the 75 source videos for evaluation. The eight lists are generated by the following schemes.

(1) *MSN*. The recommendation results from MSN Video [2011]. This is used as the baseline.
(2) *V + A* (*Visual and Aural relevances*). Using the linear combination of visual and aural features with a set of predefined weights, without considering the textual relevance and visual concepts (described in Section 4.2).
(3) *T* (*Textual relevance*). Using the linear combination of textual features with a set of predefined weights, without considering the visual and aural relevances.
(4) *MR-* (*Multimodal Relevance*). Using the linear combination of textual, visual, and aural relevances with a set of predefined weights, except for the automatic detection of visual concepts (described in Section 4.2).
(5) *MR+* (*Multimodal Relevance*). Using the linear combination of all the textual, visual, and aural relevances with a set of predefined weights, including the visual concepts.
(6) *AFF* (*Attention Fusion Function*). Fusing textual, visual, and aural relevances by AFF with a set of predefined weights.
(7) *AFF + RF* (*Attention Fusion Function and Relevance Feedback*). Fusing textual, visual, and aural relevances by AFF and RF.
(8) *AFF + RF + VT* (*Attention Fusion Function, Relevance Feedback, and shot feature adjustment based on voting approach, described in Section 5.1*). Fusing textual, visual, and aural relevances by AFF, RF, and the voting-based approach.

Please note that schemes (1)–(4) are the same as the settings in our previous work [Mei et al. 2007c; Yang et al. 2007], while the other schemes are proposed in this article. We will show in the next sections that schemes (5)–(8) outperform the previous efforts. The predefined weights used in schemes (2)–(6) are listed in Table IV. Obviously, such an empirical setting could not satisfy all kinds of videos owing to their diverse characteristics. However, we will show that our approach is able to automatically adjust these weights for different videos. The weights in Table IV were only used as the initial values. Since it is difficult to objectively evaluate the relevance of two video documents, we conducted a subjective user study. We invited 20 evaluators majoring in computer science, including 10 graduate and 10 undergraduate students. For each video in the selected 75 source videos, the subject was asked to first browse the source video and get familiar with the content. Then, the subject was provided with the recommended videos returned by the eight schemes (in a mixed manner) in a random order. The subjects did not know by which scheme the current video was recommended. After viewing these videos, they were asked to give a rating score for each recommended video (1–5), indicating whether the recommended videos are relevant to current videos and their interests (higher score indicating higher relevance).

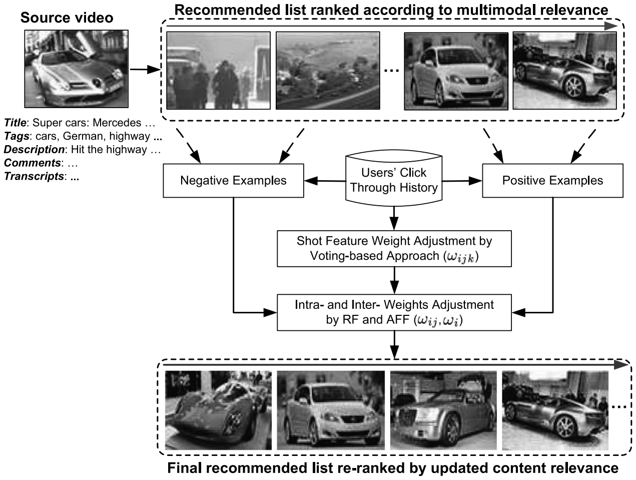

Source video
Recommended list ranked according to multimodal relevance
...
Title: Super cars: Mercedes …
Tags: cars, German, highway ...
Description: Hit the highway …
Comments: …
Transcripts: ...
Negative Examples
Users' Click Through History
Positive Examples
Shot Feature Weight Adjustment by Voting-based Approach (ω_{ijk})
Intra- and Inter- Weights Adjustment by RF and AFF (ω_{ij}, ω_i)
...
Final recommended list re-ranked by updated content relevance

Fig. 5. Procedure of the proposed approach in schemes (7) and (8).

Each subject was assigned 12 source videos so that each source video and its returned recommendations were evaluated at three times ($3.2 = \frac{12\times20}{75}$).

In schemes (7) and (8), the videos with the average rating score higher than "3" were regarded as "positive," while the videos with average rating score lower than "2" were "negative." Therefore, the intra- and interweights are adjusted according to these examples. The procedure of the proposed approach in schemes (7) and (8) is shown in Figure 5. For a given source video, we first generate a recommended list to a user according to current intra- and interweights. Then, from this user's click-through, we classify some videos in the current video list into "positive" or "negative" examples, and update the historical "positive" and "negative" lists which were obtained from the user's previous click-through. Finally, the intra- and interweights are updated based on the new "positive" and "negative" lists, and are used for the next user. At the same time, the weights of the features in each shot are automatically adjusted according to users' browsing behaviors described in Table III. Since we only have 13,000 videos in total, the respond time for recommending top 10 videos for a given video is less than 0.5 sec. Note that all the multimodal features are processed offline. If the number of videos is huge (e.g., millions of videos in the database), we can practically use the textual relevance in g_3 to obtain an initial short list of recommended videos (e.g., the top 1,000 candidate videos for recommendation), and then use the proposed approach to obtain the final ranked list of recommended videos.

6.2. Evaluation of Multimodal Relevance

To evaluate the effectiveness of different modalities, we first compared the performances of schemes (1)–(5). Similar to traditional recommendation and search system, we use the Average Rating score (AR), Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG) which are computed over the top 1, 5, and 10 recommended videos as the performance metrics [Adomavicius and Tuzhilin 2005; Liu et al. 2009]. AR is computed among the ratings from users on all the videos, while MAP is the mean of noninterpolated Average Precisions (AP). The videos with rating scores higher than 3 are defined as relevant documents while computing AP. In summary, AR indicates the average rating of all videos, while MAP indicates the ranking order of the "correct" recommendations in the list. NDCG is a commonly adopted

Table V. Performance of Five Schemes in Terms of AR

	MSN	V+A	T	MR-	MR+	AFF	AFF+RF	AFF+RF+VT
Top 1	2.90	1.40	3.25	3.30	3.45	3.15	3.42	3.55
Top 5	2.67	1.00	3.09	3.10	3.15	2.86	3.31	3.40
Top 10	2.65	1.03	2.88	2.82	2.85	2.80	3.16	3.26

Table VI. Performance of Five Schemes in Terms of MAP

	MSN	V+A	T	MR-	MR+	AFF	AFF+RF	AFF+RF+VT
Top 1	0.54	0.006	0.66	0.68	0.70	0.68	0.74	0.79
Top 5	0.48	0.010	0.61	0.62	0.64	0.58	0.69	0.76
Top 10	0.49	0.013	0.58	0.59	0.61	0.53	0.65	0.70

Table VII. Performance of Five Schemes in Terms of NDCG@d

	MSN	V+A	T	MR-	MR+	AFF	AFF+RF	AFF+RF+VT
$d = 1$	0.61	0.30	0.67	0.72	0.76	0.70	0.78	0.82
$d = 5$	0.53	0.24	0.62	0.69	0.71	0.68	0.74	0.78
$d = 10$	0.51	0.16	0.55	0.60	0.64	0.60	0.70	0.75

metric for evaluating a search engine's performance. Given a query, the NDCG score at the depth d (i.e., the top d documents) in the ranked documents is defined by

$$NDCG@d = Z_d \sum\nolimits_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1 + j)}, \tag{15}$$

where r^j is the rating of the j-th document, Z_d is a normalization constant and is chosen so that a perfect ranking's $NDCG@d$ value is 1. The results are listed in Tables V–VII.

We can see from the results that the performances of "V+A" are relatively low. From the AR, MAP, and NDCG in this scheme, we can see that only a few videos were correctly recommended. This is because that the low-level visual and aural features cannot well present the relevance at semantic level without textual information. Even though two videos are quite similar in terms of visual and aural features, their contents are probably not relevant at all. The results also show that the scheme "T" using textual information is better than the baseline (i.e., MSN Video [2011]). Moreover, the results from the linear combination of all kinds of relevance (i.e., "MR-") are better than those from pure textual information on average. This indicates that the visual and aural features can improve the performance of traditional video recommendation. However, since we only use the predefined weights for all videos, the improvements are not so significant. We also observed that by leveraging visual concepts (i.e., automatically recognized categories used in "MR+"), we can achieve much better performance in terms of multimodal relevance.

6.3. Evaluation of Relevance Fusion

From the discussion in Section 6.2, we can observe that scheme (5) (MR+) achieves the best performance among the five linear combination scheme (1)–(5). Further, we integrated the fusion strategies of AFF and RF into MR+ for evaluation. The results are listed in Table V–VII. We can see that the performance of AFF is better than that of linear combination. When using RF, the performance improves significantly in terms of AR, MAP, and NDCG in the top 1, 5, and 10 videos. Furthermore, the improvement will be more significant with the increase of users and their browsing behaviors

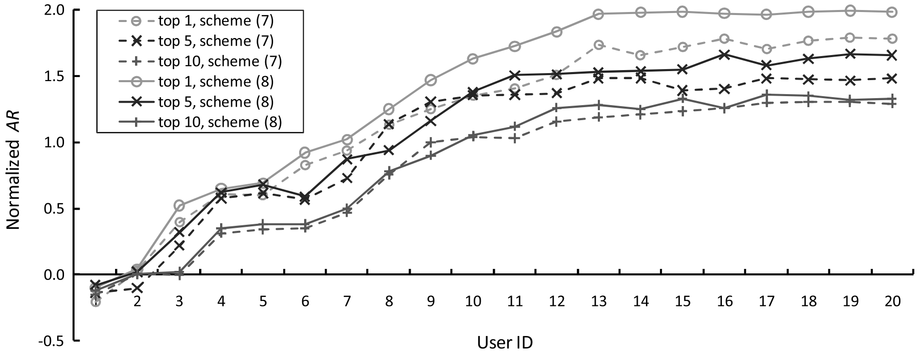


Fig. 6. Variation of the normalized AR by schemes (7) and (8). Please note that the dashed lines indicate the results from scheme (7), while the solid lines indicate the results from scheme (8).

when VT is integrated. From the performance of AFF+RF+VT, we can conclude that leveraging users' browsing behaviors on certain video segments can help video recommendation.

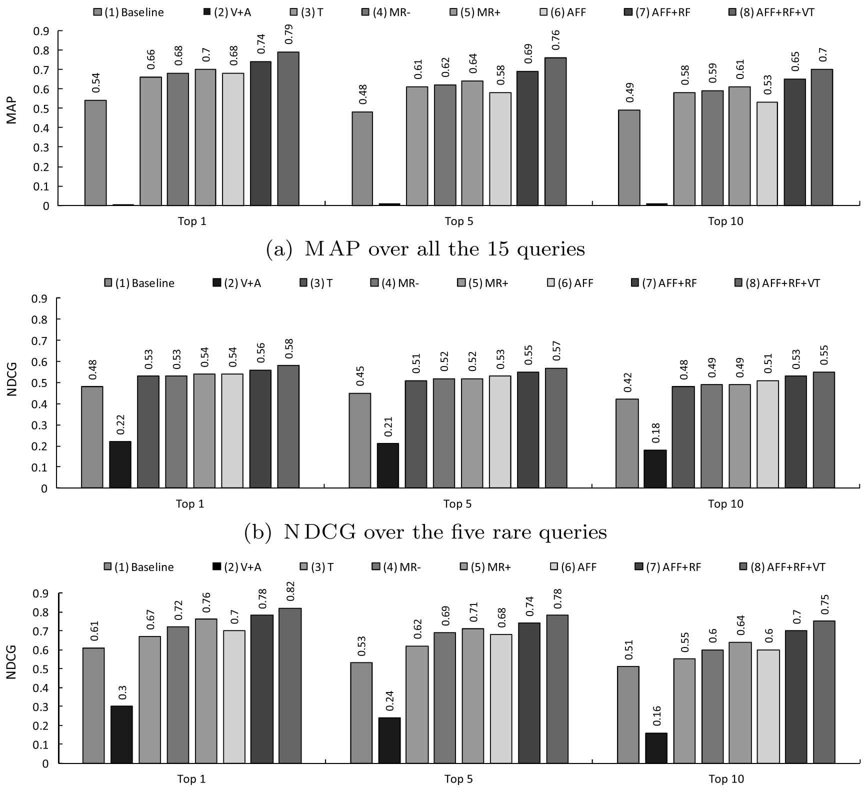
To see the effectiveness of weights adjustment using AFF+RF+VT, we traced the variation of AR with the increase of users. Since different users may have different measurements during the evaluation, we use *normalized AR* as the comparative satisfaction score among users. The normalized AR is defined as the individual rating of a single user from the corresponding scheme minus his/her average rating of the videos from all the eight schemes. Here, we use the ratings of all other schemes as a baseline. If the normalized AR is above zero, then it indicates that the corresponding scheme improves the performance. For example, if a user's rating toward scheme (8) is 4.5, while his average rating among all schemes is 3.0, then his normalized AR is 1.5 = 4.5 − 3.0. By adopting normalized AR, we significantly reduce the bias caused by different scoring strictness of different users. Since it is difficult to normalize MAP and NDCG, we only use normalized AR here. The normalized average scores of top 1, 5, and 10 of different users are shown in Figure 6. The users are sorted by order of participation from earliest to latest.

From Figure 6, we summarize the conclusions as follows.

- The overall positive slope indicates continually improved performance. The performance increases when the number of users increases, which indicates the effectiveness of relevance feedback.
- Most of the normalized AR are above zero, which indicates scheme AFF+RF+VT outperforms the other schemes.
- The normalized AR of top 1 is higher than that of top 5 and 10 for most users, which indicates users are more satisfied with top 1 videos than others. Therefore, the most relevant videos have been pushed in top recommendation list.
- Scheme (8) outperforms scheme (7), which indicates that integrating the voting-based approach for adjusting the weights of different shots can improve the recommendation performance.

We can compare the performances among the eight schemes in Figure 7. It is observed that by AFF+RF+VT, VideoReach achieved consistent improvements in top 1, 5, and 10 recommended videos. Although the MAP and NDCG are not significantly improved by AFF, both AR and NDCG have been improved by AFF.

From Figure 7(a), we can see that the average improvement of MAP from schemes (1)–(3) is 22.56%, while that from schemes (3)–(8) is 21.66%. This improvement in



(a) MAP over all the 15 queries

(b) NDCG over the five rare queries

(c) NDCG over all the 15 queries

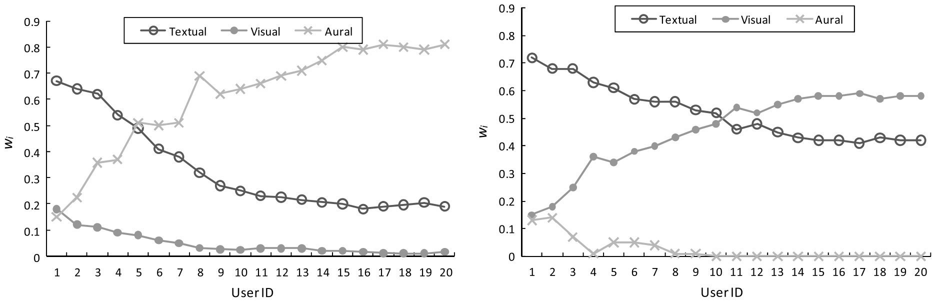
Fig. 7. Comparison among the eight schemes in terms of MAP and NDCG.

terms of NDCG from schemes (1)–(3) is 11.55%, while that from schemes (3)–(8) is 28.19%, which can be observed from Figure 7(c). These observations indicate that although the textual relevance in scheme (3) can improve the baseline relevance, relevance feedback and the voting-based approach can further significantly improve the textual relevance, and thus improve the overall recommendation performance.

We are also interested in investigating the different performance on popular and rare queries. Figure 7(b) shows the NDCG over the five rare queries. We can see that compared with the average NDCG over all queries, those on the rare queries are lower. Moreover, compared with the significant improvements over all queries (as we discussed in the preceding paragraph), the improvement over the rare five queries is 12.68% from schemes (1)–(3), 7.97% from schemes (3)–(7), and 3.66% from schemes (7)–(8), respectively. These observations indicate that our approach is more suitable for popular queries than rare queries. This is partially because rare queries usually return less relevant videos and have less specific semantics.

6.4. Evaluation of Interweights Adjustment

To show the adjustment of interweights for different videos, we use the video of “flower” and “beach” in Figure 8 as examples. It is observed that for the video of “flower,” the weight of aural relevance increases significantly in terms of the number of users, while the weight of visual relevance decreases gradually. It is reasonable that the most important characteristics of a music clip about “flower” predominantly come from aural



(a) weights adjustment of the "flower" video (b) weights adjustment of the "beach" video

Fig. 8. Adjustments of interweights for different videos.

modality, while for the video about "beach," the weight of visual relevance increased while that of aural relevance decreased. This indicates that the visual modality is able to describe "beach" better than the others. These observations indicate that different modalities tend to have different levels of importance for different queries, in other words, the interweights are query dependent. The interweight tends to be highly related to the discriminative capability of the corresponding modality; the more discriminative the modality, the higher the corresponding interweight. From Figure 8(a), we can see that the aural modality is more discriminative than visual and textual modalities, which results in the higher interweight of aural modality. By contrast, for the "beach" query, visual and textual modalities are able to differentiate the videos, leading to the high interweights of these two modalities in Figure 8(b).

The contribution from the relevance feedback (for adjusting intra- and interweights) is significant. This can be derived from Figure 7. The improvement from scheme (3) (i.e., only textual relevance) to scheme (7) (i.e., leveraging the relevance feedback technique) in terms of MAP is 12.43%, while that in terms of NDCG is 21.01%.

In real cases, since there are large numbers of users whose click-through can be used for the adjustment of weights, the proposed system will achieve better performance with sufficient "positive" and "negative" examples. To deal with continuous increase of users' click-through, we can only save most recent click-through, which may better represent current public views on a video. We admit that sometimes the performance may decrease a little with a new incoming noise click-through. However, noises do not very strongly coincide as most public click-through does. Therefore, as long as the pool of recent click-through is big enough, the overall performance will become better.

7. CONCLUSIONS AND FUTURE WORK

In this article, we have proposed a novel video-driven recommender called VideoReach, which is able to recommend a list of the most relevant videos according to a user's current viewing without his/her profile. We describe the relevance between two video documents from multimodality. We have shown how relevance feedback is leveraged to automatically adjust the intraweights within each modality and interweights between different modalities based on user click-through. Furthermore, we fuse the relevances from different modalities using an attention fusion function to exploit the variance of relevances among different modalities. The experiments indicate the effectiveness of VideoReach for online recommendation of video content. VideoReach is an effective complement to current collaborative recommendation systems which are predominantly driven by a large collection of user profiles.

Video has become one of the most compelling aspects of online media properties. With the right strategy and the right technology for recommendation, we can leverage video content for a more effective recommendation. In this work, we have gone one step further from existing online recommendation and proposed VideoReach as one of the first attempts towards contextual video recommendation. This work has shown how video and audio content analysis can yield more contextual video recommendation by effectively leveraging existing techniques for video and audio analysis.

We believe that there is rich potential for future research on video recommendation. From the perspective of contextual relevance to video content and user interest, the future work includes: (1) using high-level visual concepts which are built only based on visual content to better describe video content [Mei et al. 2007a], so that VideoReach can deal with the videos with very poor tags (as we have mentioned in Section 4.2); (2) supporting more elaborate recommendation based on video shots instead of the whole video, so that we can build a dynamic recommender which can vary the recommended video lists according to the current varying video content; (3) and collecting user profiles (such as user behaviors [Hu et al. 2007] and location [Kennedy et al. 2008]) from the click-through to make the recommendation more targeted to the user.

ACKNOWLEDGMENTS

We would like to express our appreciation to all the participants in user study experiments, and thank L. Ma and L. Yang for building some modules in this system. We also thank the anonymous reviewers for their valuable comments.

REFERENCES

ADOMAVICIUS, G. AND TUZHILIN, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Engin. 17*, 6, 734–749.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison Wesley.

BALABANOVIC, M. 1998. Exploring versus exploiting when learning user models for text recommendation. *User Model. User-Adapt. Interact. 8*, 4, 71–102.

BALUJA, S., SETH, R., SIVAKUMAR, D., ET AL. 2008. Video suggestion and discovery for youtube, taking random walks through the view graph. In *Proceedings of the International World Wide Web Conference*.

BOLL, S. 2007. Multitube-Where multimedia and web 2.0 could meet. *IEEE Multimedia Mag. 14*, 1, 9–13.

BOLLEN, J., NELSON, M. L., ARAUJO, R., AND GEISLER, G. 2005. Video recommendations for the open video project. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*. 369–369.

BURKE, R. 2002. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact. 12*, 4, 331–370.

CHANG, S.-F., MA, W.-Y., AND SMEULDERS, A. 2007. Recent advances and challenges of semantic image/video search. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

CHRISTAKOU, C. AND STAFYLOPATIS, A. 2005. A hybrid movie recommender system based on neural networks. In *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40*, 65.

ENCYCLOPEDIA. 2011. Encyclopedia. http://www.encyclopedia.com/.

FOUSS, F., PIROTTE, A., RENDERS, J. M., AND SAERENS, M. 2007. Random-Walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Trans. Knowl. Data Engin. 19*, 3, 355–369.

GIBAS, M., CANAHUATE, G., AND FERHATOSMANOGLU, H. 2008. Online index recommendations for high-dimensional databases using query workloads. *IEEE Trans. Knowl. Data Engin. 20*, 2, 246–260.

GU, Z., MEI, T., HUA, X.-S., TANG, J., AND WU, X. 2008. Multi-Layer multi-instance learning for video concept detection. *IEEE Trans. Multimedia 10*, 8, 1605–1616.

HAUPTMANN, A. G., CHRISTEL, M. G., AND YAN, R. 2008. Video retrieval based on semantic concepts. *Proc. IEEE 96*, 4, 602–622.

HU, J., ZENG, H.-J., LI, H., NIU, C., AND CHEN, Z. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the International World Wide Web Conference*.

HUA, X.-S., LU, L., AND ZHANG, H.-J. 2004a. Optimization-Based automated home video editing system. *IEEE Trans. Circ. Syst. Video Tech. 14*, 5, 572–583.

HUA, X.-S. AND ZHANG, H.-J. 2004b. An attention-based decision fusion scheme for multimedia information retrieval. In *Proceedings of the IEEE Pacific-Rim Conference on Multimedia*.

IWATA, T., SAITO, K., AND YAMADA, T. 2008. Recommendation method for improving customer lifetime value. *IEEE Trans. Knowl. Data Engin. 20*, 9, 1254–1263.

KENNEDY, L., CHANG, S.-F., AND NATSEV, A. 2008. Query-Adaptive fusion for multimodal search. *Proc. IEEE 96*, 4, 567–588.

LEW, M. S., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-Based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Comm. Appl. 2*, 1, 1–19.

LIU, Y., MEI, T., AND HUA, X.-S. 2009. CrowdReranking: Exploring multiple search engines for visual search reranking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 500–507.

MEI, T., HUA, X.-S., LAI, W., YANG, L., ET AL. 2007a. MSRA-USTC-SJTU at TRECVID 2007: High-Level feature extraction and search. In *Proceedings of TREC Video Retrieval Evaluation Online*.

MEI, T., HUA, X.-S., YANG, L., AND LI, S. 2007b. VideoSense: Towards effective online video advertising. In *Proceedings of ACM Multimedia*. 1075–1084.

MEI, T., YANG, B., HUA, X.-S., YANG, L., YANG, S.-Q., AND LI, S. 2007c. VideoReach: An online video recommendation system. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*. 767–768.

MOXLEY, E., MEI, T., AND MANJUNATH, B. S. 2010. Video annotation through search and graph reinforcement mining. *IEEE Trans. Multimedia 12*, 3, 184–193.

MSN VIDEO. 2011. MSN video. http://video.msn.com/video.aspx?mkt=en-us&tab=soapbox/.

NAPHADE, M., SMITH, J. R., TESIC, J., CHANG, S.-F., HSU, W., KENNEDY, L., HAUPTMANN, A., AND CURTIS, J. 2006. Large-Scale concept ontology for multimedia. *IEEE Multimedia Mag. 13*, 3, 86–91.

RESNICK, P. AND VARIAN, H. R. 1997. Recommender systems. *Comm. ACM 40*, 3, 56–58.

RUI, Y., HUANG, T. S., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circ. Video Tech. 8*, 5, 644–655.

SETTEN, M. V. AND VEENSTRA, M. 2003. Prediction strategies in a TV recommender system—Method and experiments. In *Proceedings of the International World Wide Web Conference*.

SHEN, D., PAN, R., SUN, J.-T., PAN, J. J., WU, K., YIN, J., AND YANG, Q. 2006a. Query enrichment for web-query classification. *ACM Trans. Inf. Syst. 24*, 3, 320–352.

SHEN, D., SUN, J.-T., YANG, Q., AND CHEN, Z. 2006b. Building bridges for web query classification. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 131–138.

SHEN, J., SHEPHERD, J., CUI, B., AND TAN, K.-L. 2009. A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inf. Syst. 27*, 3.

SHEN, J., TAO, D., AND LI, X. 2008. Modality mixture projections for semantic video event detection. *IEEE Trans. Circ. Syst. Video Tech. 18*, 11, 1587–1596.

SIERSDORFER, S., PEDRO, J. S., AND SANDERSON, M. 2009. Automatic video tagging using content redundancy. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–402.

SNOEK, C. G. M. AND WORRING, M. 2009. Concept-based video retrieval. *Found. Trends Inf. Retr. 4*, 2, 215–322.

SNOEK, C., WORRING, M., VAN GEMERT, J., GEUSEBROEK, J.-M., AND SMEULDERS, A. W. M. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*. 421–430.

TAO, D., TANG, X., LI, X., AND WU, X. 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Patt. Anal. Mach. Intell. 28*, 7, 1088–1099.

TRECVID. 2011. TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

WEI, Y. Z., MOREAU, L., AND JENNINGS, N. R. 2005. Learning users interests by quality classification in market-based recommender systems. *IEEE Trans. Knowl. Data Engin. 17*, 12, 1678–1688.

YAHOO! 2011. Yahoo. http://www.yahoo.com/.

YANG, B., MEI, T., HUA, X.-S., YANG, L., YANG, S.-Q., AND LI, M. 2007. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. 73–80.

YANG, Y. AND LIU, X. 1999. A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

YOUTUBE. 2011. YouTube. http://www.youtube.com/.

YU, B., MA, W.-Y., NAHRSTEDT, K., AND ZHANG, H.-J. 2003. Video summarization based on user log enhanced link analysis. In *Proceedings of the ACM International Conference on Multimedia*. 382–391.

ZHOU, D., ZHU, S., YU, K., SONG, X., TSENG, B. L., ZHA, H., AND GILES, C. L. 2008. Learning multiple graphs for document recommendations. In *Proceedings of the International World Wide Web Conference*. 141–150.

Received January 2010; revised August 2010; accepted December 2010