Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback*

Bo Yang †, Tao Mei ‡, Xian-Sheng Hua ‡, Linjun Yang ‡, Shi-Qiang Yang †, Mingjing Li ‡

† Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China
‡ Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P. R. China
bo.yang02@gmail.com; {tmei, xshua, linjuny, mjli}@microsoft.com; yangshq@mail.tsinghua.edu.cn

ABSTRACT

With Internet delivery of video content surging to an unprecedented level, video recommendation has become a very popular online service. The capability of recommending relevant videos to targeted users can alleviate users' efforts on finding the most relevant content according to their current viewings or preferences. This paper presents a novel online video recommendation system based on multimodal fusion and relevance feedback. Given an online video document, which usually consists of video content and related information (such as query, title, tags, and surroundings), video recommendation is formulated as finding a list of the most relevant videos in terms of multimodal relevance. We express the multimodal relevance between two video documents as the combination of textual, visual, and aural relevance. Furthermore, since different video documents have different weights of the relevance for three modalities, we adopt relevance feedback to automatically adjust intra-weights within each modality and inter-weights among different modalities by users' click-though data, as well as attention fusion function to fuse multimodal relevance together. Unlike traditional recommenders in which a sufficient collection of users' profiles is assumed available, this proposed system is able to recommend videos without users' profiles. We conducted an extensive experiment on 20 videos searched by top 10 representative queries from more than $13k$ online videos, reported the effectiveness of our video recommendation system.

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*

General Terms

Algorithms, Human Factors, Experimentation.

Keywords

online video recommendation, multimodal fusion, relevance feedback

1. INTRODUCTION

Driven by the age of Internet generation and the advent of near-ubiquitous broadband Internet access, online delivery of video content have surged to an unprecedented level. According to an Online Publishers Association study [17], more than 140 million people (69%) have watched video online with 50 million (24%) doing so weekly. This trend has brought a variety of online video services, such as video search, video tagging and editing, video sharing, video advertising, and so on. Therefore, it is natural to imagine that today's online users always face a daunting volume of video content - be it from video sharing or blog content, or from IPTV and mobile TV. As a result, there is an increasing demand of an online video service to push the "interesting" or "relevant" content to targeted people at every opportunity. Video recommendation is such a kind of service which releases users' efforts on manually filtering out the unrelated content and finding the most interesting videos according to their current viewings or preferences. While many existing video-oriented sites, such as YouTube [6], MySpace [5], Yahoo! [4], Google Video [2] and MSN Soapbox [1], have already provided recommendation services, it is likely that most of them recommend the relevant videos only based on surrounding text information (such as the title, tags, and comments). However, it still remains a challenging research problem to leverage video content and users' click-though data for a more efficient recommendation.

The earlier research on recommendation began with Resnick *et al.*, who has given a general definition for a recommender system as to assist and augment the natural social process [18]. A typical recommender system receives the recommendations provided by users as inputs, and then aggregates and directs to appropriate recipients aiming at good matches between recommended items and users. While in the specific domain of online video service, the input of a video recommendation system is the video content clicked by a user, together with related information (such as query and surrounding text provided by content providers), and the output is a list of recommended videos according to user's current views and preference (such as user interest and location).

*This work was performed while the first author was visiting Microsoft Research Asia as a research intern.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

There exists rich research on video recommendation. Most of the previous work on traditional recommendation returned a personalized list of videos based on users' profile, with the assumption that a sufficient collection of users' profiles is available. However, in many real cases, a user visits a web page anonymously and is less likely to login the system to provide his/her personal profile. Thus, traditional recommendation approaches cannot be directly applied to current online video recommendation. An alternative to video recommendation is to adopt the techniques used in video search. However, the tasks of video search and video recommendation are quite different. Video search aims at finding the videos that mostly "match" a query. In other words, those videos "relevant" to the query instead of directly "matching" the query will not be returned in a video search system. While in a video recommendation system, both the "matching" and "relevant" videos are desirable to be recommended to the users. In addition, the input query in a search system is usually a list of keywords or together with an example of image, while in a recommendation system it consists of not only keywords but also video clicked by a specific user and surrounding text related to this video. Therefore, multimodal relevance should be taken into account for video recommendation.

Motivated by these observations, we propose a novel online video recommendation system using multimodal relevance between two video documents and users' click-through data. In our system, an online video is represented as a video document, which includes not only video (such as visual and aural content, as well as ASR/OCR embedded in video stream), but also related information (such as the query, title, tags, and surroundings). Given a video document, which is selected by a user, the recommender aims at finding a list of the most relevant videos. We believe that the relevance between 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 attention fusion function (AFF) successfully used in multimedia information retrieval by exploiting the different variance among multimodal relevance. Furthermore, relevance feedback is used to automatically adjust intra-weights within each modality and inter-weights among different modalities by using users' click through data. The experiments have indicated that our system outperforms MSN Soapbox, and showed how the performance can be improved by relevance feedback with users' click-through data as well.

The rest of this paper is organized as follows: Section 2 reviews previous work related to video recommendation; our system framework is presented 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 gives experimental results, followed by conclusions in Section 7.

2. RELATED WORK

The research problems closely related to our work are traditional recommendation and video search. We will briefly review previous work on these two issues, and describe the essential differences to our work in the next.

2.1 Traditional Recommendation

Research on traditional recommendation started from 1990s. Many recommendation systems have been designed in diverse areas, such as movies [9] [16], TVs [20], web pages [8], and so on. It is observed that most of these recommenders assumed that a sufficient collection of users' profiles is available. In general, users' profiles mainly come from two kinds of sources: (1) direct profiles, i.e. users' selection of a list of predefined interests; (2) indirect profiles, i.e. users' ratings of a number of items. To summarize, regardless of what kinds of items were recommended by these systems, the objective is to recommend the items matching to uses' profiles. In other words, the "relevance" in traditional recommendation systems is based on users' profiles or interests instead of the item itself.

However, while surfing web, most users usually browse web page anonymously. Thus, in many cases, an online video recommender has to deal with the absence of users' profiles. Therefore, traditional recommending techniques cannot be directly applied in online video scenario.

2.2 Video Search

The techniques used in video search can be classified into two categories: text-based and content-based.

Text-based video search is to search videos according to surrounding texts, while content based search is to leverage visual content for searching. Recently, a great deal of effort has been carried out on content based video search [10] [11] [15], where content features can be used directly to compute the similarities between videos, or used with users' interactive evaluations, or used for re-ranking the results returned by text based search.

However, video recommendation is quite different from video search, which makes it unsuitable to directly adapt the techniques used in video search to video recommendation. We summarize the differences listed as follows:

- Different objectives. Video search is to find videos that mostly "match" the queries or a query image in concept level. However, the objective of video recommendation is to rank the videos which are most "relevant" with a source video, as we have discussed in Section 1.
- Different inputs. The input of video search comes from a set of keywords or images, where all inputs do not have any property; while the input of video recommendation consists of video and its related surroundings. Moreover, the surroundings usually have specific properties, such as query, title, tags, comments, and so on. Therefore, it is desirable to deal with the different importance of multimodal information.

3. SYSTEM FRAMEWORK

The input of our video recommendation system is a video document $\mathbf{D}$, which is represented by textual, visual and aural documents as $\mathbf{D} = (D_T, D_V, D_A)$. Given an online video document $\mathbf{D}$ clicked by a user, the task of video recommendation is expressed as finding a list of videos with the best relevance to $\mathbf{D}$. Since different modalities have different contributions to the relevance, we use $(\omega_T, \omega_V, \omega_A)$ to denote

the weights of textual, visual and aural document, respectively. Thus a video document can be further represented by

$$\mathbf{D} = \mathbf{D}(D_T, D_V, D_A, \omega_T, \omega_V, \omega_A) \quad (1)$$

Similarly, the document of a single modal D_i $(i \in \{T, V, A\})$ can be represented by a set of features and the corresponding weights:

$$D_i = D_i(\mathbf{f}_i, \omega_i) \quad (2)$$

where $\mathbf{f}_i = (f_{i1}, f_{i2}, \ldots, f_{in})$ is a set of features from modality i, and $\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{in})$ is a set of corresponding weights. Let $\mathcal{R}(D_x, D_y)$ denotes the relevance of two video documents D_x and D_y. The relevance between video document D_x and D_y in terms of modality i is denoted by $\mathcal{R}_i(D_x, D_y)$, while the relevance in terms of feature f_{ij} is denoted by $R_{ij}(D_x, D_y)$.

Figure 1 illustrates system framework of our online video recommendation. To obtain multimodal relevance between two video documents, the relevance in terms of a single modality is first computed by weighted linear combinations of relevance between features. Then relevance of single modality is fused using attention fusion function (AFF) with the weights proposed in [14]. The intra-weights within each modality and inter-weights among different modalities are adjusted dynamically using relevance feedback [19].

Using textual features to compute the relevance of video documents is the most common method and can work well in most cases. However, not all concepts can be well described by text only. For instance, 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 unrelated videos, such as "desert", "weather", and so on. In this case, it is better to use visual features to describe "beach" rather than text. Furthermore, aural features are quite important for relevance in some music videos. Therefore, in addition to textual features, we use visual and aural features to augment the description of all types of online videos. We next describe the relevance from textual, visual and aural documents, as well as fusion strategy by AFF and relevance feedback.

4. MULTIMODAL RELEVANCE

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

4.1 Textual Relevance

We classify textual information related to a video document into two kinds: (1) *direct* text, referring to the surrounding text provided by users themselves, closed captions (CC), Automated Speech Recognition (ASR), and Optical Character Recognition (OCR) embedded in video stream; (2) *indirect* text, referring to the categories and their probabilities obtained by automatic text categorization based on a set of predefined category hierarchy. We use the vector and probabilistic models to describe direct and indirect text, respectively. Thus a textual document D_T is represented as

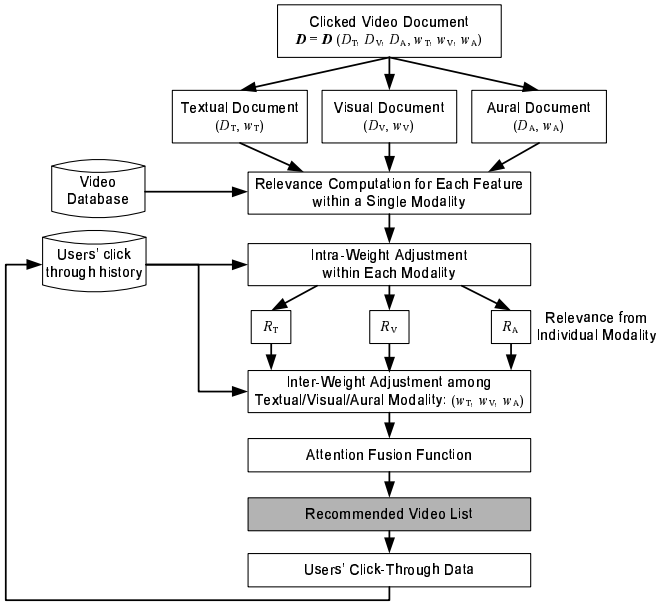

Figure 1: System framework of proposed online video recommendation.

two kinds of features (f_{T1}, f_{T2}) from vector and probabilistic models

$$D_T = D_T(f_{T1}, f_{T2}, \omega_{T1}, \omega_{T2}) \quad (3)$$

where w_{T1} and w_{T2} indicate the weights of f_{T1} and f_{T2}, respectively.

4.1.1 Vector Model

In the vector model, the textual feature of a document is usually defined as

$$f_{T1} = f_{T1}(\mathbf{k}, \omega) \quad (4)$$

where $\mathbf{k} = (k_1, k_2, \ldots, k_n)$ is a dictionary of all keywords appearing in the whole document pool, $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ is a set of corresponding weights, n is the number of unique keywords in all documents.

A classic algorithm to calculate the importance of a keyword is to use the product of its *term frequency* (TF) and *inverted document frequency* (IDF), based on the assumption that the more frequently a word appears in a document and the rarer the word appears in all documents, the more informative it is. However, such approach is not suitable in our scenario. First, the number of keywords related to online videos is smaller than that of a text document, which leads to a small document frequencies (DF). Thus, IDF is quite unstable according to its common definitions [7]. Second, most online content providers tend to use general keywords to describe their videos, such as "car" instead of "Benz." Using IDF will make some non-informative keywords overwhelm the informative ones. Therefore, we only use "tf" to describe the importance of a keyword. In the vector model, *cosine* distance is adopted as the measurement of textual relevance between document D_x and D_y

$$\mathcal{R}_{T1}(D_x, D_y) = \frac{\omega(D_x) \cdot \omega(D_y)}{|\omega(D_x)||\omega(D_y)|} \quad (5)$$

where $\omega(D_x)$ denote the weights of D_x. Different kinds of text may have different weights. The more a kind of text is related with the video document, the more important it is. Since the title and tags provided by content providers are usually more relevant to uploaded videos than the other textual information, their corresponding weights are set higher (i.e. 1.0); while the weights of comments, descriptions, ASR, and OCR are lower (i.e. 0.1).

4.1.2 Probabilistic Model

Although vector model is able to present the keywords of a textual document, it is not enough to describe the 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 vector model. Consequently, many videos related to real flowers will be recommended by the vector model. However, the videos related to music are more relevant. To address this problem, we leverage the category information obtained by probabilistic model. We use text categorization based on Support Vector Machine (SVM) [21] to automatically classify a textual document into a set of predefined category hierarchy which consists of more than $1k$ categories.

In our probability model, textual feature of D_T is represented as

$$f_{T2} = f_{T2}(\mathbf{C}, \mathbf{P}) \tag{6}$$

where $\mathbf{C} = (C_1, C_2, \ldots, C_m)$ is a set of categories to which the textual document D_T is belonging with a set of probabilities $\mathbf{P} = (P_1, P_2, \ldots, P_m)$.

The predefined categories make up a hierarchical category tree. Let $\mathrm{d}(C_i)$ denote the depth of category C_i in category tree, where the depth of root is 0. For two categories C_i and C_j, we define $\ell(C_i, C_j)$ as the depth of their first common ancestor. Then for two textual documents D_x, with a set of categories $C_x = (C_1, C_2, \ldots, C_{m1})$ and probabilities $P_x = (P_1, P_2, \ldots, P_{m1})$, and D_y with $C_y = (C_1, C_2, \ldots, C_{m2})$ and $P_y = (P_1, P_2, \ldots, P_{m2})$, the relevance in probabilistic model is defined as

$$\mathcal{R}_{T2}(D_x, D_y) = \sum_{i=1}^{m1} \sum_{j=1}^{m2} \mathcal{R}(C_i, C_j) \tag{7}$$

where $\mathcal{R}(C_i, C_j) = \alpha^{(d(C_i)-\ell(C_i,C_j))} P_i \cdot \alpha^{(d(C_j)-\ell(C_i,C_j))} P_j$, if $\ell(C_i, C_j) > 0$; otherwise, $\mathcal{R}(C_i, C_j) = 0$. α is a predefined parameter controlling the probabilities of upper-level categories. Intuitively, the deeper level two documents are similar at, the more related they are. For instance, in Figure 2, for the two nodes B and C represented by (C_i, P_i) and (C_j, P_j), thus $\mathcal{R}(C_i, C_j) = \alpha^2 P_i \cdot \alpha^3 P_j = \alpha^5 P_i P_j$. In our experiments, α is fixed to 0.5.

4.2 Visual Relevance

The visual relevance is measured by normalized color histogram, motion intensity and shot frequency (the average number of shots per second), which had proved to be effective to describe visual content in many existing video retrieval systems [12] [13].

A visual document D_V is represented as

$$D_V = D_V(f_{V1}, f_{V2}, f_{V3}, \omega_{V1}, \omega_{V2}, \omega_{V3}) \tag{8}$$

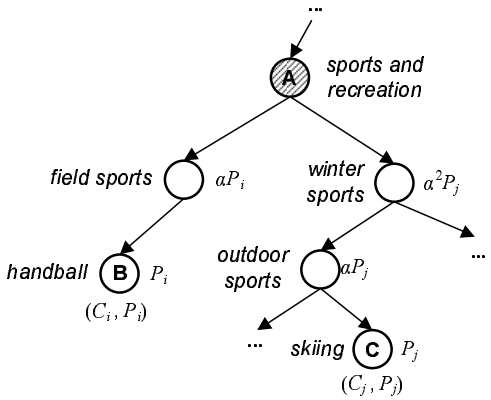

Figure 2: The hierarchical category tree. Each node denotes a category. A is the first common ancestor of B and C.

where f_{V1}, f_{V2} and f_{V3} represent color histogram, motion intensity, and shot frequency, respectively. For two visual documents D_x and D_y, the visual relevance of feature j $(j = 1, 2, 3)$ is defined as

$$\mathcal{R}_{Vj}(D_x, D_y) = 1.0 - |f_{Vj}(D_x) - f_{Vj}(D_y)| \tag{9}$$

4.3 Aural Relevance

As for an aural document, we describe it using the average and standard deviation of aural tempos among all the shots. Average aural tempo represents the speed of music or audio, while standard deviation indicates the change frequency of music style. These features have proved to be effective to describe aural content [12].

As a result, an aural document D_A is represented as

$$D_A = D_A(f_{A1}, f_{A2}, \omega_{A1}, \omega_{A2}) \tag{10}$$

where f_{A1} and f_{A2} represent the average and standard deviation of aural tempo, respectively. For two aural documents D_x and D_y, the aural relevance of these features is defined as

$$\mathcal{R}_{A1}(D_x, D_y) = 1.0 - |f_{A1}(D_x) - f_{A1}(D_y)| \tag{11}$$
$$\mathcal{R}_{A2}(D_x, D_y) = 1.0 - |f_{A2}(D_x) - f_{A2}(D_y)| \tag{12}$$

5. FUSION STRATEGY

We have modeled the relevance from individual channels. However, fusing these relevancies to a final measurement for recommendation is another key issue. We will show how to combine the relevance from individual modality by attention fusion function and relevance feedback.

5.1 Fusion with Attention Fusion Function

Linear combination of the relevance of individual modality is a straightforward and effective method for fusion. However, this approach is not consistent with human's attention response. Hua *et al.* have proposed an Attention Fusion Function (AFF) to simulate human's attention characteristics [14]. 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 present high relevance in one individual modality but low relevance in the other, they still have a high final relevance.

In our system, the first property is easy to be satisfied in our scenario since each component contributes to the overall relevance; while for the second, since two documents are not necessarily relevant even they are very similar in terms of one feature, we first fuse the above relevance into three channels: textual, visual, and aural relevance. If two documents have high textual relevance, they are probably relevant; if two documents are only quite similar in visual or aural features, they are still not very relevant. Therefore, we first filter out most documents in terms of textual relevance to assure all documents are more or less relevant with the input document, and then only calculate the visual and aural relevance within these documents. Thus, if a document has high visual or aural relevance with the clicked video, user will pay more attention to it than to others with all moderate relevance scores according to the attention model.

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 AFF with weights in [14] to get a final relevance. For two documents D_x and D_y, the final relevance is computed as

$$\mathcal{R}(D_x, D_y) = \frac{R_{avg} + \frac{1}{2(n-1)+n\gamma}\sum_i |n\omega_i \mathcal{R}_i(D_x, D_y) - R_{avg}|}{W} \quad (13)$$

where

$$R_{avg} = \sum_i \omega_i \mathcal{R}_i(D_x, D_y) \quad (14)$$

$$W = 1 + \frac{1}{2(n-1)+n\gamma}\sum_i |1 - n\omega_i| \quad (15)$$

and $i \in \{T, V, A\}$. n is the number of modalities ($n = 3$), ω_i is the weight of individual modality to be detailed at next section, γ is a predefined constant and fixed to 0.2 in our experiments. For more details of AFF function, please refer to [14].

5.2 Adjust Weights with Relevance Feedback

Before using AFF to fuse relevance from three modalities, two issues need to be addressed: (1) how to obtain the intra-weights of relevance for each kind of feature within a single modality (e.g. ω_{T1} and ω_{T2} in textual modality); (2) how to decide the inter-weights (i.e. ω_T, ω_V and ω_A) of relevance for each modality.

Actually, it is hard 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 important. Therefore, it is better to assign different video documents with different intra- and inter-weights. It is observed that users' click-through data usually tell a latent instruction to the assignment of weights, or at least a latent comment on the recommendation results. 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.

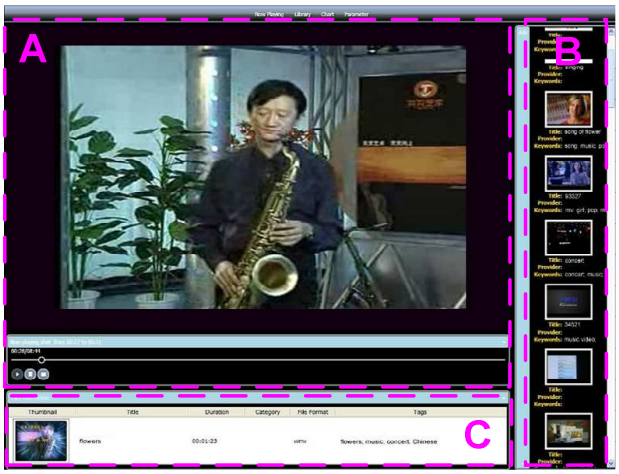

Figure 3: User Interface of our video recommendation system. A – online video; B – recommended video list; C – related textual descriptions of this online video.

However, if a user views a recommended video for a relative long time, this video is probably a true recommendation, since this user is rather interested in this recommendation. We call such videos "positive" examples. With "positive" and "negative" examples, relevance feedback [19] is an effective solution to automatically adjusting the weights of different inputs, i.e. intra- and inter- weights.

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

$$\omega_{ij} = \frac{1}{\sigma_{ij}} \quad (16)$$

where $i \in \{T, V, A\}$, σ_{ij} is the standard deviation of feature f_{ij}, whose corresponding document D_i is a positive example. The intra-weights are then normalized between 0 and 1.

The adjustment of inter-weights is to obtain the optimal weight of each modality. For each modality, a recommendation list $(D_1, D_2, \ldots, D_K)$ is created based on the individual relevance from this modality, where K is the number of recommended videos. We first initialize $\omega_i = 0$, and then update w_T as follows

$$\omega_i = \begin{cases} \omega_i + 1, & \text{if } D_k \text{ is a "positive" example} \\ \omega_i - 1, & \text{if } D_k \text{ is a "negative" example} \end{cases}$$

where $i \in \{T, V, A\}$ and $k = 1, \ldots, K$. The inter-weights are then normalized between 0 and 1.

6. EXPERIMENTS

6.1 Experimental Setup

We have collected more than $13k$ online videos from MSN Soapbox [1] in our video database. It is not reasonable to evaluating our system over all these source videos. Instead, we use 20 representative source videos for evaluation. These videos are searched by 10 popular queries from our database. The content of these videos covered a diversity

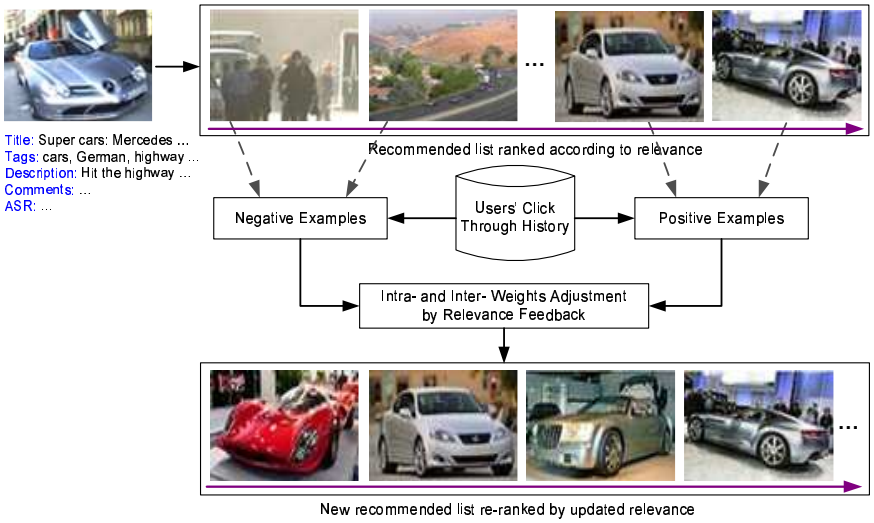

Figure 4: Procedure of our proposed approach.

of genres, such as music, sports, cartoon, movie previews, persons, travel, business, food, and so on. The selected 10 representative queries came from the most popular queries in MSN [3] excluding sensitive and similar queries. These queries include "flowers," "cat," "baby," "sun," "soccer," "fire," "beach," "food," "car," and "Microsoft." We input these queries our system to search a list of videos, in which only top two are selected as source videos in our evaluation. Figure 3 shows user interface of our video recommendation system. For each source video, we recommended six different lists of videos, with each containing 20 videos. The six lists are generated by the following schemes:

1. Soapbox. The recommendation results from "MSN Soapbox" [1], as our baseline.

2. VA (Visual+Aural Relevance). Using the linear combination of visual and aural features with predefined weights.

3. Text (Textual Relevance). Using linear combination of textual features with predefined weights.

4. MR (Multimodal Relevance). Using linear combination of textual, visual and aural information with predefined weights.

5. AFF (Attention Fusion Function). Fusing textual, visual and aural information by AFF with predefined weights.

6. AFF + RF (AFF + Relevance Feedback). Using textual, visual and aural information with relevance feedback and attention fusion function.

The predefined weights used in scheme 2∼5 are listed in Table 1. Since it is difficult to evaluate the relevance of two videos documents objectively, we conducted a subjective user study. We invited 10 evaluators majored in computer science, including eight graduate students and two

Table 1: Predefined weights in scheme 2∼5

	ω_T		ω_V			ω_A	
weight	ω_{T1}	ω_{T2}	ω_{V1}	ω_{V2}	ω_{V3}	ω_{A1}	ω_{A2}
intra	0.5	0.5	0.5	0.3	0.2	0.7	0.3
inter	0.70		0.15			0.15	

undergraduate students. Each individual is given the recommended videos returned by the six approaches in a random order. After viewing these videos, they were asked to give a rating score from 1 to 5 indicating whether the recommended videos are relevant to current viewings (higher score indicating more relevance).

In AFF+RF, those videos with average rating score bigger than 3 are regarded as "positive" examples and the ones with average rating score smaller than 2 as "negative" examples. Thus, intra- and inter- weights are adjusted according to these examples. The procedure of our approach is shown in Figure 4. For an input video document, we first generate a recommended list to a user according to current intra- and inter- weights; then from this user's click-through, we classify some videos in the list into "positive" or "negative" examples, and update the historical "positive" and "negative" lists which are obtained from previous users' click-through; finally, the intra- and inter- weights are updated based on the new "positive" and "negative" lists, and are used for the next user.

6.2 Results of Linear Combination

In order to see the effectiveness of different modalities, we first compare the performances of scheme 1∼4. Similar to traditional recommendation and search system, we use the average rating score (AR), average accuracy (AC) and mean average precision (MAP) of top 5, 10, and 20 recommended videos as the measurements. AC is defined as the proportions of videos with the rating bigger than 4 to all recommended videos. The mean average precision is the mean of non-interpolated average precisions (AP) [13]. The videos with scores no less than 4 are defined as relevant documents

Table 2: Average ratings (AR) of four schemes

	Soapbox	VA	Text	MR
Top 5	2.6719	1.0000	2.9906	3.0063
Top 10	2.6625	1.0320	2.8078	2.7695
Top 20	2.6172	1.0172	2.6395	2.7777

Table 3: Average accuracies (AC) of four schemes

	Soapbox	VA	Text	MR
Top 5	0.3038	0.0000	0.3488	0.3438
Top 10	0.2894	0.0050	0.2950	0.3006
Top 20	0.2738	0.0025	0.2766	0.2825

Table 4: Mean average precisions (MAP) of four schemes

	Soapbox	VA	Text	MR
Top 5	0.4618	0.0000	0.5944	0.6147
Top 10	0.4830	0.0125	0.5705	0.5876
Top 20	0.4499	0.0125	0.5089	0.5166

when computing AP. In summary, AR indicates the average rating of all videos, AC indicates the proportion of correct recommendations, and MAP indicates the ranking order of correct recommendations in the list. The results are listed in Table 2, 3 and 4.

The results show that the performances of VA are relatively low. From the AR, AC and MAP in VA scheme, we can see that only few videos are correctly recommended. This is because low-level visual-aural features could not present the relevance at semantic level well without textual information. Even though two videos are quite similar in terms of visual-aural features, their contents are probably not relevant at all. The results also show that our scheme using textual information (Text) is better than "Soapbox." Moreover, the results of using all kinds of information (MR) are better than those of using only textual information in average. This indicates that visual-aural features can improve the performance of recommendations. However, since we only use the predefined weights for all videos, the improvements are not so significant. We will show how to significantly improve the performance by AFF and RF in the next.

6.3 Results of AFF and RF

It is observed that scheme 4 (i.e. MR) gives the best performance among the linear combination schemes 1∼4. Thus we integrate the fusion strategy of AFF and RF into MR for evaluation. The results are listed in Table 5. From the above results, we can see that the performance of AFF is better than that of linear combination. When using relevance feedback together, the performance improves significantly, in terms of all AR, AC and MAP in top 5, 10 and 20. Furthermore, the improvement will be more significant with the increase of users and their feedback.

To see the effectiveness of weights adjustment using AFF+RF, we trace the variation of AR with the increase of users. Since different users may have different measurements during the evaluation, we use *normalized score* as comparative satisfaction among users. The *normalized score* is defined as the individual rating of a single user by AFF+RF minus his/her average rating of the videos by all the six schemes except for VA. Since almost all the ratings by VA scheme are identical (i.e. 1), these ratings can not represent users' measurements. Therefore, we do not include the ratings by VA when computing the average rating score in order to represent users' strictness more precisely. Since it is difficult to normalize AC and MAP, we only use normalized AR here. The normalized average scores of top 5, 10, and 20 of different users are shown in Figure 5. The users are sorted by the order of participation from the earliest to latest.

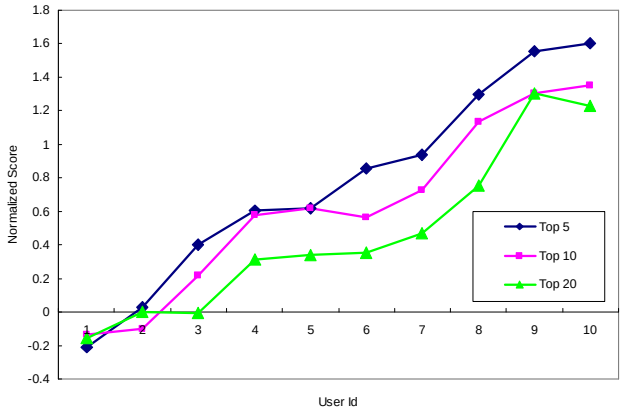

Figure 5: Variation of normalized AR by AFF + RF.

Table 5: Performances of AFF and AFF+RF

	AFF			AFF + RF		
	AR	AC	MAP	AR	AC	MAP
Top 5	3.0328	0.3500	0.6198	3.2547	0.4013	0.6836
Top 10	2.7977	0.3031	0.5876	3.1383	0.3838	0.6537
Top 20	2.7914	0.2878	0.5169	2.9461	0.3388	0.6193

From Figure 5, we summarize the conclusions as follows:

- The performance increases when the number of users increases, which indicates the effectiveness of relevance feedback.
- Most of the *normalized scores* are above zero, which indicates AFF+RF outperforming the other schemes.
- The n*ormalized scores* of top 5 is higher than that of top 10, and 20 for most users, which indicates the most relevant videos being pushed in the front of recommendation list.

In real cases, since there are large amounts of users, whose click-through can be used for the adjustment of weights, our proposed system will achieve better performance with sufficient positive and negative examples.

7. CONCLUSIONS

In this paper, we have proposed a novel online video recommendation system that is able to recommend a list of

most relevant videos according to a user's current viewing without his/her profile. We describe the relevance of two video documents from textual, visual and aural modality. We have shown how relevance feedback is leveraged to automatically adjust the intra-weights within each modality and inter-weights between modalities based on users' click-through data. Furthermore, we fuse the relevance from different modalities using attention fusion function to exploit the variance of relevance among different modalities. We conducted an extensive experiment using 20 source videos as users' current viewings, which are searched by 10 representative queries from an existing popular video site. The comparisons have indicated the effectiveness of our system for online recommendation of video content.

The future work would be mainly focused on three parts: using high-level feature extraction [13] to better describe video content, supporting more precise recommendation based on video shots instead of the whole video, and collecting user profiles (such as user interest and location) from click-through data to improve the performance.

8. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 60573167 and the National High-Tech Research and Development Plan of China (863) under Grant No. 2006AA01Z118. We would like to express our appreciation to all the participants in user study experiments.

9. REFERENCES

[1] http://soapbox.msn.com/.
[2] http://video.google.com/.
[3] http://video.msn.com/.
[4] http://video.yahoo.com/.
[5] http://www.myspace.com/.
[6] http://www.youtube.com/.
[7] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
[8] M. Balabanovic. Exploring versus exploiting when learning user models for text recommendation. *User Modeling and User-Adapted Interaction*, 8(4):71–102, Nov 1998.
[9] C. Christakou and A. Stafylopatis. A hybrid movie recommender system based on neural networks. In *Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications*, Wroclaw, Poland, 2005.
[10] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen. Extreme video retrieval: Joint maximization of human and computer performance. In *Proceedings of the ACM International Conference on Multimedia*, Santa Barbara, USA, 2006.
[11] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the ACM International Conference on Multimedia*, Santa Barbara, USA, 2006.
[12] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Trans. on Circuit and System for Video Technology*, 14(5):572–583, May 2004.
[13] X.-S. Hua, T. Mei, W. Lai, and *et al.* Microsoft Research Asia TRECVID 2006 high-level feature extraction and rushes exploitation. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.
[14] X.-S. Hua and H.-J. Zhang. An attention-based decision fusion scheme for multimedia information retrieval. In *Proceedings of IEEE Pacific-Rim Conference On Multimedia*, Tokyo, Japan, 2004.
[15] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. on Multimedia Computing, Communications and Applications*, 2(1):1–19, Feb 2006.
[16] H. Mak, I. Koprinska, and J. Poon. INTIMATE: A web-based movie recommender using text categorization. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, Beijing, China, 2003.
[17] Online Publishers. http://www.online-publishers.org/.
[18] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, May 1997.
[19] Y. Rui, T. S. Huang, and M. Ortega. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on Circuit and System for Video Technology*, 8(5):644–655, Sep 1998.
[20] M. V. Setten and M. Veenstra. Prediction strategies in a TV recommender system – method and experiments. In *Proceedings of International World Wide Web Conference*, Budapest, Hungary, 2003.
[21] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval*, California, USA, 1999.