VideoReach: an online video recommendation system.

Conference Paper · January 2007

DOI: 10.1145/1277741.1277899 · Source: DBLP

6 authors, including:

Tao Mei
Microsoft
393 PUBLICATIONS **8,560** CITATIONS

SEE PROFILE

Xian-Sheng Hua
Microsoft
332 PUBLICATIONS **9,342** CITATIONS

SEE PROFILE

Shipeng Li
Cogobuy Group & IngDan Technology
369 PUBLICATIONS **8,059** CITATIONS

SEE PROFILE

VideoReach: An Online Video Recommendation System

Tao Mei †, Bo Yang ‡, Xian-Sheng Hua †, Linjun Yang †, Shi-Qiang Yang ‡, Shipeng Li †
† Microsoft Research Asia, Beijing 100080, P. R. China
‡ Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China
{tmei, xshua, linjuny, spli}@microsoft.com; bo.yang02@gmail.com

ABSTRACT

This paper presents a novel online video recommendation system called *VideoReach*, which alleviates users' efforts on finding the most relevant videos according to current viewings without a sufficient collection of user profiles as required in traditional recommenders. In this system, video recommendation is formulated as finding a list of relevant videos in terms of multimodal relevance (i.e. textual, visual, and aural relevance) and user click-through. Since different videos have different intra-weights of relevance within an individual modality and inter-weights among different modalities, we adopt relevance feedback to automatically find optimal weights by user click-though, as well as an attention fusion function to fuse multimodal relevance. We use 20 clips as the representative test videos, which are searched by top 10 queries from more than 13k online videos, and report superior performance compared with an existing video site.

Categories and Subject Descriptors: H.3.5 [Information Storage and Retrieval]: Online Information Services – *Web-based services*

General Terms: Algorithms, Human Factors, Experimentation.

Keywords: video recommendation, multimodal relevance.

1. INTRODUCTION

Online video services have surged to an unprecedented level in recent years. Today's online users always face a daunting volume of video content from video sharing and blog sites, or from IPTV and mobile TV. As a result, there is an increasing demand of an online video recommendation system to find the most relevant videos according to users' current viewings or preferences. While many existing video-oriented sites, such as YouTube [9], MSN Soapbox [5], Yahoo! [7], and Google Video [1], have already provided recommendation services, it is likely that most of them recommend videos only based on surrounding text. However, it remains a challenging problem to leverage video content and user click-through for a more effective recommendation.

Most of previous work on traditional recommendation focus on collaborate filtering based on a sufficient collection of user profiles, e.g. the famous movie recommender system – "moviefinder" [4]. However, there are many cases that a user tends to visit a web page anonymously without providing his/her profile. Thus, traditional recommendation approaches are not appropriate for online video recommendation. Another similar area to video recommendation is video search. However, both the task and input of video search are quite different from those of video recommendation. Video search aims at finding videos that mostly "match" a set of keywords or/and a sample image; while it is important for a recommender to find the most relevant videos according to all kinds of video-related information. For example, "apple" is quite relevant with, but does not match, "orange".

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

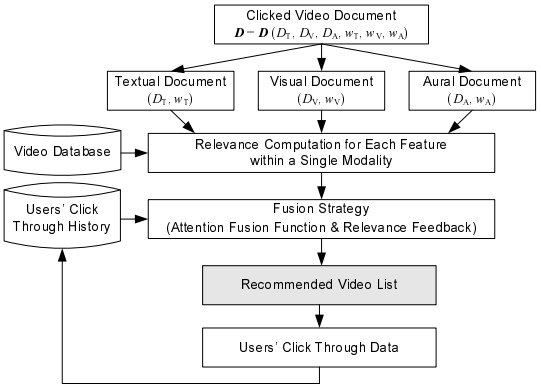

Figure 1: System Framework of VideoReach.

Motivated by the above observations, we propose *VideoReach*, a novel online video recommendation system using multimodal relevance between two video documents and user click-through simultaneously. The input of VideoReach is a video document, represented by textual, visual, and aural document. 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 from different features. Then the relevance from textual, visual and aural modalities is fused using *attention fusion function* [3] that is successfully applied in multimedia information retrieval. The intra-weights within each modality and inter-weights among different modalities are dynamically adjusted by *relevance feedback* [6] using user click-through. Figure 1 shows overview of VideoReach.

2. ONLINE VIDEO RECOMMENDATION

2.1 Multimodal Relevance

Using textual features to compute the relevance of video documents is the most common method. In VideoReach, textual information is classified into the following two types: (1) *direct* text – referring to query, title, and keywords pro-

vided by users, and embedded text such as closed captions in video stream; (2) *indirect* text – referring to the categories obtained by automatic text categorization based on a set of predefined categories. We use the vector and probabilistic models to describe the *direct* and *indirect* text, respectively.

In vector model, a textual document is represented as a set of keywords and corresponding weights. We use tf instead of classic $tf \cdot idf$ to represent the weight for each keyword, and adopt *cosine* distance to measure the relevance of two documents. We further introduce probabilistic model to describe latent semantics. Support Vector Machine based text categorization [8] is adopted to automatically classify a textual document into a set of predefined hierarchy that consists of more than $1k$ categories. For two textual documents D_x with a set of categories $C_x = (C_1, C_2, \ldots, C_{m_x})$ and the corresponding probabilities $P_x = (P_1, P_2, \ldots, P_{m_x})$, and D_y with $C_y = (C_1, C_2, \ldots, C_{m_y})$ and $P_y = (P_1, P_2, \ldots, P_{m_y})$, the relevance in probabilistic model is defined as $\sum_{i=1}^{m_x} \sum_{j=1}^{m_y} \mathcal{R}(C_i, C_j)$, where

$$\mathcal{R}(C_i, C_j) = \alpha^{(d(C_i)-\ell(C_i,C_j))} P_i \cdot \alpha^{(d(C_j)-\ell(C_i,C_j))} P_j \quad (1)$$

and $d(C_i)$ denotes the depth of category C_i in the category tree, $\ell(C_i, C_j)$ denotes the depth of the first common ancestor of C_i and C_j, α is a predefined parameter.

The visual information is described by color histogram, motion intensity and shot frequency (i.e. average number of shots per second), while the aural information is described by the mean and standard variation of aural tempo among all the shots. These features have proved to be effective to describe video content [2]. The relevance in terms of feature i between documents D_x and D_y is defined as $1.0 - |f_i(D_x) - f_i(D_y)|$, where $f_i(D_x)$ is i-th feature of D_x.

2.2 Fusion Strategy

In order to fuse the relevance from three modalities, we adopt three dimensional AFF [3], which simulates human attention nature. We first use linear combination to fuse the relevance from different features within each single modality, and then fuse the relevance from three modality using AFF by considering both *monotonicity* and *heterogeneity* simultaneously. For more details, please refer to [3].

Since different videos have different characteristics, it is difficult to select a set of weights satisfying all kinds of videos. We adopt RF [6] to automatically adjust the intra- and inter- weights for each video. We get "positive" and "negative" examples from user click-through. If a user opens a recommended video and closes it within a short time (e.g. less than five seconds), it is taken as a "negative" example; if a user views a recommended video for a relative long time, it is taken as a "positive" example. Given "positive" and "negative" examples, RF automatically adjusts the weights.

3. EXPERIMENTS

We used a collection of more than $13k$ online videos from "MSN Soapbox" [5]. Since it is not reasonable to evaluating our system over all these videos, we use 20 representative videos which are searched by 10 popular queries from our database. For each video, we recommended six different lists of videos, each containing 20 videos. The six lists are generated by the following schemes:

1. Soapbox. The recommendation results from "MSN Soapbox" [5], as our baseline.

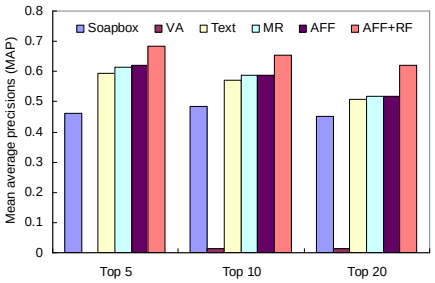

Figure 2: Mean average precisions of sixe schemes.

2. VA (Visual + Aural Relevance). Using 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. Fusing textual, visual and aural information by AFF with predefined weights.
6. AFF+RF. Using textual, visual and aural information with RF and AFF.

It is difficult to evaluate the relevance of two video documents objectively. Instead, we conducted a subjective user study. We invited 10 individuals to evaluate all recommended videos by VideoReach (i.e. give a rating score from 1 to 5, where higher score indicating more relevance) returned by the six schemes in a random order. We adopted mean average precision (MAP) of top 5, 10, and 20 recommended videos as the measurements. The videos with scores no less than 4 are defined as relevant documents when computing MAP. The results are listed in Fig. 2. It is observed that a significant improvement is obtained by using AFF and RF.

4. CONCLUSIONS

This paper presented a novel online video recommender – *VideoReach*, which is able to recommend a list of the most relevant videos according to user current viewing and click-through. We describe the relevance between two online video documents in terms of multimodal relevance. Relevance feedback is leveraged to automatically adjust intra-weights within each modality and inter-weights between modalities based on user click-through, as well as attention fusion function is used to fuse multimodal relevance.

5. REFERENCES

[1] Google Video. http://video.google.com/.

[2] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Trans. on CSVT*, 14(5):572–583, May 2004.

[3] X.-S. Hua and H.-J. Zhang. An attention-based decision fusion scheme for multimedia information retrieval. In *Proceedings of PCM*, 2004.

[4] MovieFinder. http://www.moviefinder.com/.

[5] MSN Soapbox. http://soapbox.msn.com/.

[6] Y. Rui, T. S. Huang, and M. Ortega. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on CSVT*, 8(5):644–655, Sep 1998.

[7] Yahoo! Video. http://video.yahoo.com/.

[8] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of ACM SIGIR*, 1999.

[9] Youtube. http://www.youtube.com/.