World Wide Web
DOI 10.1007/s11280-011-0129-1

Boosting web video categorization with contextual information from social web

Xiao Wu · Chong-Wah Ngo · Yi-Ming Zhu · Qiang Peng

Received: 7 September 2010 / Revised: 31 March 2011
Accepted: 31 March 2011

Abstract Web video categorization is a fundamental task for web video search. In this paper, we explore web video categorization from a new perspective, by integrating the model-based and data-driven approaches to boost the performance. The boosting comes from two aspects: one is the performance improvement for text classifiers through query expansion from related videos and user videos. The model-based classifiers are built based on the text features extracted from title and tags. Related videos and user videos act as external resources for compensating the shortcoming of the limited and noisy text features. Query expansion is adopted to reinforce the classification performance of text features through related videos and user videos. The other improvement is derived from the integration of model-based classification and data-driven majority voting from related videos and user videos. From the data-driven viewpoint, related videos and user videos are treated as sources for majority voting from the perspective of video relevance and user interest, respectively. *Semantic meaning* from text, *video relevance* from related videos, and *user interest* induced from user videos, are combined to robustly

Categories and Subject Descriptors H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval-*Relevance feedback*, *Query formulation*; H.3.5 [Information Storage and Retrieval]: On-line Information Services-*Data sharing, Web-based services*;

General Terms Algorithms, Design, Experimentation, Performance.

X. Wu (✉) · Y.-M. Zhu · Q. Peng
Department of Computer Science and Engineering, Southwest Jiaotong University,
No. 111, North Section 1, 2nd Ring Road, Chengdu, China
e-mail: wuxiaohk@home.swjtu.edu.cn

Y.-M. Zhu
e-mail: yiming.z.jtu@gmail.com

Q. Peng
e-mail: qpeng@home.swjtu.edu.cn

C.-W. Ngo
Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue,
Kowloon, Hong Kong
e-mail: cwngo@cs.cityu.edu.hk

Published online: 15 April 2011

determine the video category. Their combination from semantics, relevance and interest further improves the performance of web video categorization. Experiments on YouTube videos demonstrate the significant improvement of the proposed approach compared to the traditional text based classifiers.

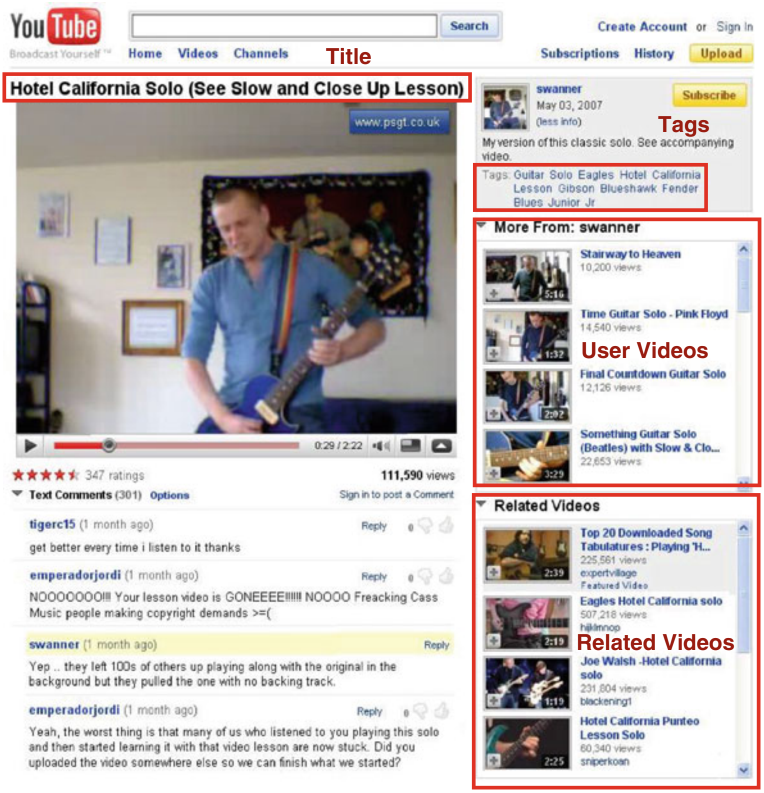
Keywords categorization · query expansion · context information · social web · web video · data-driven

1 Introduction

Web video categorization refers to the process of assigning web videos to predefined categories such as sports, news, music, and so on, which plays a vital role in many information retrieval tasks. Traditionally, video classification is undergone mainly by building a large set of classifiers on textual, audio, visual low-level features or their combination [2, 27, 28]. They heavily rely on building models by machine learning techniques (e.g., SVM, GMM) to map low-level features to high-level semantics. Although high-level concept detection techniques (e.g., [10, 15]) based on local visual keypoints (e.g., [12]) have been made significant progress recently, the performance of current high-level concept detection is far from satisfactory and the cost of feature processing is extremely expensive (usually hundreds to thousands of local points in each keyframe). The diversity of web videos ranging from professional videos to low-quality home videos makes the web video categorization more challenging. The state-of-the-art content-based video classification cannot meet the expected performance. Therefore, text information becomes the most direct, feasible and efficient feature to classify web videos. However, in social web, the number of user-supplied text information (title and tags) is limited, and usually they are noisy, ambiguous, incomplete, and even misleading. Many important terms may be absent, leading to a poor coverage of the video content. As a result, the discriminative capability of text features is deteriorated in the web scenario. The inherent ambiguity of short title and tags demands for more advanced approaches for web video categorization.

Fortunately, the social web, such as YouTube, provides rich contextual and social resources associated with videos, such as related videos, user and community information (as shown in Figure 1). The related videos frequently have relevant contents or similar category labels with the given video. At the same time, users share videos based on their personal interests, and therefore the uploaded videos by the same user usually have similar type. For example, videos from user “stanforduniversity” are associated with “Education”, while videos from user “CBS” mainly belong to “News and Politics”. These contextual resources arouse new perspectives for web video categorization. In this paper, contextual information refers to related videos, user videos and their associated text information and category label.

To overcome the shortcoming of short and noisy text features, query expansion is adopted in this paper to boost the web video classification performance by integrating information from related videos and user videos. *Query expansion*, also known as *pseudo-relevance feedback* has been proven to be effective in text information retrieval [1, 7, 11, 18, 25], which aims for improving the retrieval performance by appending additional terms to short queries. In traditional query expansion, additional query terms are extracted from the highly relevant items in the initial retrieval run, and then the expanded query is run again to return a fresh set of documents to user. The prosperity of the social web provides such a platform for query expansion. Enlightened by the idea of query expansion, in this paper, we explore techniques to boost the effectiveness of text classification for web video



Hotel California Solo (See Slow and Close Up Lesson)

★★★★☆ 347 ratings

111,590 views

Text Comments (301)

tigerc15 (1 month ago)
get better every time i listen to it thanks

emperadorjordi (1 month ago)
NOOOOOOO!!! Your lesson video is GONEEEE!!!!! NOOOO Freacking Cass Music people making copyright demands >=(

swanner (1 month ago)
Yep .. they left 100s of others up playing along with the original in the background but they pulled the one with no backing track.

emperadorjordi (1 month ago)
Yeah, the worst thing is that many of us who listened to you playing this solo and then started learning it with that video lesson are now stuck. Did you uploaded the video somewhere else so we can finish what we started?

swanner
May 03, 2007
(less info)
My version of this classic solo. See accompanying video.

Tags: Guitar Solo Eagles Hotel California Lesson Gibson Blueshawk Fender Blues Junior Jr

More From: swanner

Stairway to Heaven
10,200 views

Time Guitar Solo - Pink Floyd
14,540 views

Final Countdown Guitar Solo
12,126 views

Something Guitar Solo (Beatles) with Slow & Clo...
22,653 views

Related Videos

Top 20 Downloaded Song Tabulatures : Playing 'H...
225,561 views
expertvillage
Featured Video

Eagles Hotel California solo
507,218 views
hijklmnop

Joe Walsh -Hotel California solo
231,804 views
blackening1

Hotel California Punteo Lesson Solo
60,340 views
sniperkoan

Figure 1 Contextual information (title, tags, related videos, user videos, and so on) associated with videos.

categorization by generating additional terms from contextual information. The associated related videos and user videos, acting as external sources, are assumed to be highly relevant to the given video. The collected terms from title/tags of related/user videos are expected to contain new and useful words, which could help improve the discriminative power of short and noisy terms, and eventually benefit for the web video classification. Classifiers based on the expanded text terms are then built with Support Vector Machines (SVM) to classify videos.

Furthermore, we integrate the model-based and data-driven approaches to boost the web video categorization. The aforementioned method based on classifiers belongs to the so-called "*model-based*" approaches. On the other hand, with the explosive expansion of social web, developing "*data-driven*" approaches becomes feasible when huge amounts of web videos are freely available on web. This paves a new way for approaching traditional applications by data-driven methodology. With overwhelming amount of community-contributed video data available, many problems can be solved without the need of sophisticated algorithms. In this paper, in addition to model-based classifiers on text, we explore web video classification from data-driven perspective, by recommending categories according to the majority voting of category labels from related videos and user videos. The dominant category indicates the preference of related/user videos from a higher level viewpoint, which gives constructive clues for the web video categorization. From data-driven viewpoint, related videos and user videos are treated as sources for majority voting from the perspective of video relevance and user interest, respectively. The fusion of model-based classifiers and data-driven majority voting further improves the performance of web video categorization. From another point of view, our approach integrates sources from three different aspects: semantic meaning, video relevance, and user interest to improve the performance. The terms from title and tags give information from the semantic aspect, while related videos and user videos indicate the video relevance and user interest, respectively. Their combination could give a more comprehensive recommendation for the video category.

In this paper, the main contribution is the novel idea of boosting web video categorization with contextual information from social web. The boosting mainly comes from two aspects. One is the performance improvement of model-based text classifiers through query expansion. In order to compensate the shortcoming of the limited and noisy text features, related/user videos act as external resources for query expansion. The classifiers are then built based on the expanded features. The other improvement is derived from the combination of model-based classification and data-driven majority voting from related/user videos. With the abundant amounts of web videos and their associated metadata, the data-driven approach becomes feasible. Semantic meaning, video relevance, and user interest are integrated from different perspectives to provide more accurate indication of video category. Another critical property to mention is that the proposed approach has high scalability. The classification based on text features and the data-driven majority voting are cheap to implement, which is especially suitable for web-scale video applications.

The rest of this paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 elaborates the techniques for boosting web video categorization. Section 4 presents the experiments. Finally, section 5 concludes this paper.

2 Related work

2.1 Query expansion

User queries tend to be short and abbreviated. To improve the retrieval performance, a wide range of methods for enriching query terms have been proposed in information retrieval, from manual techniques such as iterative query refinement and relevance feedback, to automatic techniques such as thesauri expansion, and query expansion [11, 25]. Relevance feedback was proposed decades ago as a method for improving the effectiveness of information retrieval. Originally, a list of documents is presented to the user, from which the user manually marks them as relevant or irrelevant. It is observed that terms closely related to relevant documents have better discriminative property to separate relevant from irrelevant documents. Additional terms extracted from relevant documents are then appended to the original query to form new query. Experiments show that the new query could significantly improve the effectiveness. However, the approach requires that users take time to assess each document. To avoid the user interaction, automatic query expansion, also called as pseudo-relevance feedback, is developed as a variant of relevance feedback. The additional query terms are extracted from top-ranked documents (for example, top 10 documents), on the assumption that they are likely to be relevant. A new query is formed based on this feedback. Approaches based on vector space model [18] and statistical language models [11, 29] have been proposed to generate the new retrieval model. Since user queries are typically short, augmenting additional relevant query terms can lead to significant improvement in retrieval effectiveness. Recently, query expansion has been extensively studied from different aspects (e.g., selecting good expansion terms [4], negative relevance feedback [23]), and been applied to various applications in text information retrieval, including mining user logs [9], sponsored search [3], federated search [20], and so on.

Relevance feedback and query expansion have been applied to multimedia community. Relevance feedback is first introduced into interactive content-based image retrieval [19]. Negative pseudo-relevance feedback based on a statistical model is proposed in [26] to improve the search performance for content-based video retrieval. Recently, semantic concept-based query expansion and re-ranking is explored in [14, 16]. Text query expansion based on lexical and statistical approaches, as well as visual query expansion based on content analysis is developed for concept-based query expansion [14]. An automatic query expansion is proposed based on a semantic concept feature space [16]. Although query expansion has been demonstrated an effective approach for text and multimedia information retrieval, few studies have examined whether query expansion is indeed helpful for web video categorization. The social web provides a source of query expansion. The useful resources are related videos and user videos, which are assumed to be relevant to the given video. Our work follows the strategy of query expansion by inducing extra terms from related/user videos and adjusting the original term weights to boost the classification performance. As far as we know, there is little work investigating the impact for web video categorization by expanding terms from related videos and user videos.

2.2 Model-based vs. data-driven

The existing approaches for video classification are dominated by model-based approaches, which depend heavily on the machine learning techniques (e.g., SVM) to build classifiers on textual, audio, visual low-level features and their combination [2, 27, 28]. The work [27] might be the most comprehensive one for web video categorization, which considers semantic modality (concept histogram and visual word representation) and surrounding text modality. The experiments demonstrate that the multimodality feature representation is effective, and SVM outperforms GMM and MR (Manifold Ranking) for all modalities. A consensus learning approach [17] is proposed to train classifiers and clustering algorithms on text, audio and video, and then combine them for multi-label web video classification. A hierarchical video genre ontology is constructed and hierarchical SVM classifiers are designated to categorize video genres [28]. However, collecting a large set of noise-free training examples with sufficient positive samples for learning is always not easy. Manual annotation of training examples is laborious. The extraction and selection of low-level visual features (e.g., global features versus local features) remains an open issue. The significant diversity of visual appearance makes the learnt models unreliable and hardly generalizable. The situation is even worse for noisy web videos. Due to the aforementioned reasons, these model-based approaches have their limitations when facing ever-increasing web videos.

On the other hand, with the huge volume of social data, data-driven approaches become a new trend for web applications. Recently, such data-driven techniques have been evident by various purposes, for instance, word similarity measure [8], object recognition [22], and image/video annotation [13, 21, 24]. Successful examples of data-driven approaches include the following applications. Normalized Google distance [8] measures the similarity between two words by querying the number of web pages containing the words from Google search engine. A novel work on “80 million tiny images” approaches the object and scene recognition from the data-driven viewpoint [22]. When the underlying images are huge enough, simple scheme like nearest neighbor matching can perform reasonably well. With the similar idea, annotating images is performed first by discovering visually and semantically similar search results, and then identifying salient terms from search results under a corpus with 2.4 million images in [24]. Such techniques, also referred to as "annotation by search", have also been demonstrated in [13, 21] for video annotation. Similar videos are ranked in a multimodal search, and graph reinforcement mining is proposed for propagating tags from similar documents to query videos [13]. Although the data-driven approaches are simple, they reflect the preference of the majority of data. Therefore, it turns out to be effective when abundant data are available, especially, under the social web scenario. In this paper, we combine the model-based classifiers on text features and the data-driven majority voting from related videos and user videos to further improve the web video categorization.

3 Boosting web video categorization

In this section, the contextual information associated with videos is first introduced, followed by a brief description of the proposed framework. Query expansion for text classification and majority voting for related/user videos are elaborated in sections 3.3 and 3.4, respectively. Finally, they are combined to further improve the performance.

3.1 Contextual Information

The community-contributed social web has resulted in rich sets of contextual information associated with web videos, which are illustrated in Figure 1. In this work, the following contextual information is exploited for our web video categorization.

- **Text (title and tags)**: Text words carry semantic meaning to briefly summarize the video content. The title and tags of a web video are thus the most direct source for web video categorization. While informative, the text information for web videos is commonly short, noisy, ambiguous, and even misleading.
- **Related Videos**: They refer to the related videos associated with a given video in YouTube (see Figure 1). The related videos are usually similar or relevant videos. The majority of related videos could help to estimate the probability of the video category.
- **User Videos**: User videos are the videos uploaded by the same user. The user-uploaded videos are commonly consistent with the user's personal interests. Therefore, to a large extent, the category label can be inferred by the majority of user videos.

3.2 Framework

In this paper, we deploy the combination of model-based and data-driven approaches to classify web videos. The overall framework is depicted in Figure 2. Various contextual features (title, tags, related videos, and user videos) are exploited. For a given video, its related videos and user videos are first exploited as query expansion to boost the text classification performance. The terms derived from the related videos and user videos are utilized to adjust the term weights of the original video. Model-based classifiers are then built based on the extended text information, which is expected to reinforce the classification performance. This component contributes to the confidence from the semantic meaning viewpoint. In addition, the category labels of related videos and user videos provide constructive clues for web video categorization. The data-driven approaches derived from the majority voting of category labels from related videos and user videos are further combined to obtain a robust measurement. From another viewpoint, they constitute

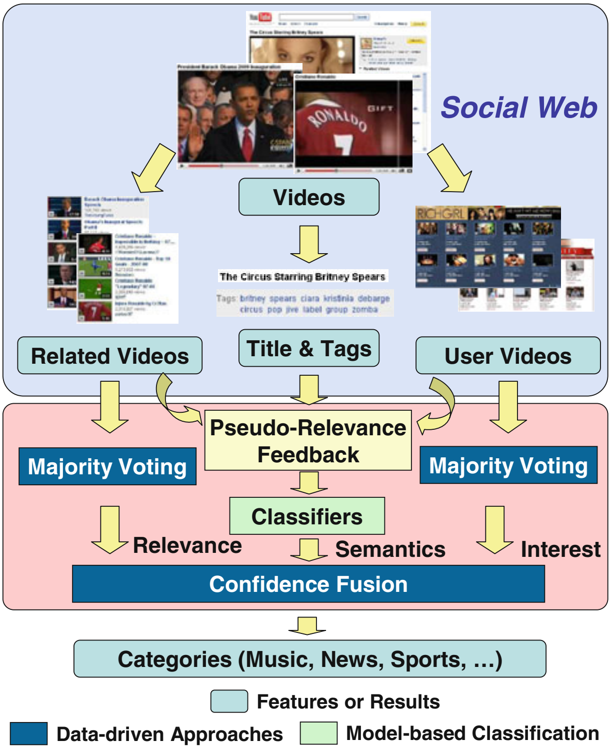

Social Web
Videos
The Circus Starring Britney Spears
Tags: britney spears ciara kristinia debarge circus pop jive label group zomba
Related Videos
Title & Tags
User Videos
Majority Voting
Pseudo-Relevance Feedback
Majority Voting
Classifiers
Relevance
Semantics
Interest
Confidence Fusion
Categories (Music, News, Sports, …)
Features or Results
Data-driven Approaches
Model-based Classification

Figure 2 The framework of web video categorization.

the confidence from video relevance and user interest, respectively. In the following sections, we will elaborate the details.

3.3 Boosting by query expansion

The text feature is one of the direct resources for web video categorization. Nevertheless, one issue is that title and tags are often too short and noisy to clearly describe the contents of videos. This problem can be compensated by introducing additional text terms induced from related videos and user videos. Related videos and user videos are assumed to be relevant to the given video, therefore, can be regarded as the source of query expansion. It is expected that the expanded terms extracted from the related/user videos can assist to discriminate the category. For example, the expanded terms, such as Bush, white house, congress, Obama and so on, will give more indication that this video is closely related to politics, and finally facilitate to classify this video as "News and Politics" category. In this section, the idea of query expansion is employed by importing additional terms from related/user videos, targeting for enhancing the discriminative power of text classifiers. Extracted representative terms are used to either append new terms or adjust the weights of existing terms. The expanded terms are then exploited to build the classifiers to categorize the videos from the perspective of semantic meaning.

Given a video, the original text terms can be represented as a feature vector. And the associated related/user videos have their corresponding title and tags, which are similarly represented as feature vectors. The related/user videos are treated as a form of query expansion. Important and frequent terms extracted from title/tags of query expansion are expected to be useful for improving the discriminative capability of different categories. A list of expanded terms, with a weight associated with each term indicating its importance, is derived from query expansion. Enlightened by the classic Rocchio algorithm [18], an expanded vector is therefore formed with a linear combination of the original vector and the new vector after expansion. The scheme for adjusting weights of query expansion is an interesting research topic to study. However, it is out of the scope of this work. We will explore this issue in our future work.

The adjusted weight of a term t_i for video V_j is determined by:

$$weight_{ij} = tf\left(w_i, V_j\right) + \alpha\left(log \sum_{V_k \in F_j} tf\left(w_i, V_k\right)\right) / |F_j|$$

where $tf(w_i, V_j)$ is the term frequency in the original vector for video V_j, F_j is the set of related/user videos (depending on strategies adopted) associated with video V_j, $|F_j|$ is the number of related/user videos, and α is a parameter used to control the weighting between the original vector and the expanding vector. The number of related videos is relatively stable, around 60 per video in YouTube. However, for different users, the number of user videos can be variance dramatically. For active users, like "CBS", the number of user videos can be as high as 27,176, while for some users, the number can be lower than 10. Therefore, the weight is normalized by the number of relevant videos $|F_j|$.

The individual classifier is then built per category based on the reweighted text vectors using SVM. For testing videos, confidence scores for categories are returned from these classifiers. $Semantics_i(V_j)$ is the confidence score of video V_j belonging to the predefined category C_i, which is the probability score returned from SVM classifier for category C_i based on text terms. The text classifiers recommend the category from the semantic view. In the following section, we will introduce the recommendation from the aspects of video relevance and user interest.

3.4 Majority voting for video relevance and user interest

One attractive aspect of community-contributed social media is the abundant amount of context metadata, which arouses new perspective for web video categorization. The related videos frequently have relevant contents or similar category labels with the given video. At the same time, users upload and share videos based on their personal interests. Therefore the uploaded videos by the same user usually have similar type. For example, videos from user "BritneyTV" are tightly associated with "Music" category, while videos from user "BMWwebTV" mainly belong to "Autos and Vehicles." By checking the category labels of related/user videos, it gives meaningful hints for the web video categorization from the data-driven aspect.

$Relevance_i(V_j)$ and $Interest_i(V_j)$ are confidence scores according to class label distribution from related videos and user videos, respectively, which are defined as follows:

$$Relevance_i(V_j) = |R_{ij}|/|R_j| \quad Interest_i(V_j) = |U_{ij}|/|U_j|$$

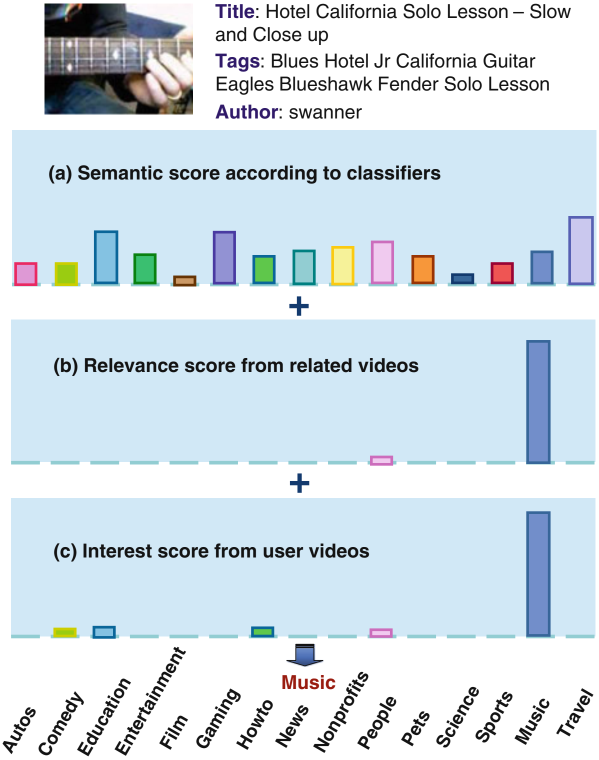
They are simple statistics of the portion of related/user videos having category label C_i. That is the percentage of videos belonging to category C_i among all the related/user videos. R_j is the set of related videos for video V_j, and R_{ij} is the set of related videos having category label C_i among R_j. Similarly, U_j is the set of user videos uploaded by the same user as video V_j, and U_{ij} is the set of user videos having category label C_i among U_j. Note that the sum of distribution is equal to 1, and m is the number of categories (in this work, m is equal to 15 based on YouTube categories).

$$\sum_{i=1}^{m} Relevance_i(V_j) = 1 \quad \sum_{i=1}^{m} Interest_i(V_j) = 1$$

A skewed category distribution with a dominant peak indicates that videos in this distribution have certain category preference. The new video has high possibility to have the same category label as the dominant category. While a flat distribution with low scores means that their relevance or interests are diverse, the incoming video has possibility to be any category. As shown in Figures 3(b) and (c), we can see that the skewed category distributions for both related videos and user videos are dominated by “Music” category. It denotes that the videos are tightly related to music, and user “swanner” has strong interest for music videos. Therefore, the new uploaded video has high potential to be with the same category.

3.5 Boosting by fusion of model-based and data-driven approaches

The text classifiers and the majority voting contribute confidence from different aspects. The semantic meaning of title and tags ($Semantics_i$), video relevance from related videos ($Relevance_i$), and user interest induced from user videos ($Interest_i$), provide complementary clues from three different viewpoints. To achieve a robust measurement, these components are integrated to further boost the performance. Since the objective is to study the effect of


Title: Hotel California Solo Lesson – Slow and Close up

Tags: Blues Hotel Jr California Guitar Eagles Blueshawk Fender Solo Lesson

Author: swanner

Figure 3 According to classifiers built on text (a), a video on guitar lesson is falsely suggested as “Travel” category. When considering video relevance (b) and user interest (c), it is correctly recommended as “Music” category.

integration of model-based and data-driven approaches, we only adopt uniform weight fusion and leave the adaptive fusion for future study. The three components are uniformly combined to give a confidence judgment *Score*$_i$(*V*$_j$), which is determined as follows:

$$Score_i(V_j) = Semantics_i(V_j) + Relevance_i(V_j) + Interest_i(V_j)$$

From the data viewpoint, this solution combines data-driven and model-based methods. Model-based classifiers are built based on text feature through query expansion, while the data-driven majority voting is evaluated based on related videos and user videos. Finally, the category having the highest score determines the category label of video V_j:

$$C_k = argmax_{i=1}^{m} Score_i(V_j)$$

4 Experiments

4.1 Dataset

To compare the performance of web video categorization, we collect 11,314 videos and their corresponding metadata (e.g., title, tags, username, youtubeid, and so on) from YouTube. Most of them are obtained from the recently released MCG_WEBV dataset [5] from ICT-CAS. The data are collected by retrieving "Most Viewed" videos of "*This Month*" (from December 2008 to July 2009) and "*All Time*" from the predefined 15 categories (such as Music, Sports, People, and so on) in YouTube. The videos from December 2008 to March 2009 are treated as the training set, while the videos of "All Time" and videos from April 2009 to July 2009 are used as the testing datasets *TS_A* and *TS_B*, respectively. Each category contains around 100 videos per month except "Music" category. Furthermore, the metadata of the related videos and videos uploaded by the same user are collected. Because part of the related/user videos are absent in MCG_WEBV dataset, we further supplement them by parsing the YouTube webpage and calling YouTube API to make the dataset as complete as possible. The detailed information on dataset is listed in Table 1. For example, there are totally 111,462 related videos, and 136,542 user videos associated with videos in TS_A. After a serial of data preprocessing (e.g. stop word removal, special character removal), there are 9,831, and 18,889 unique words in TS_A and TS_B respectively. The original video category labels are treated as the ground truth.

4.2 Performance metrics

We use *precision* to evaluate the performance, which is defined as:

$$Precision_i = \left|P_i^+\right| / |P_i|$$

Table 1 Dataset Information.

Dataset	Core Videos	Related Videos	User Videos	Unique Words
Training	4,610	199,063	258,547	20,577
TS_A	2,047	111,462	136,542	9,831
TS_B	4,657	234,859	126,522	18,889

where P_i^+ is the number of correctly classified positive samples for category C_i, and P_i is the number of positive samples in ground truth. And the *average precision* (AP) is adopted to measure the overall performance for m categories (m=15):

$$AP = \left(\sum_{i=1}^{m} Precision_i\right)/m$$

4.3 Performance comparison for model-based and data-driven methods

Firstly, we compare the performance of model-based classification and data-driven majority voting for web video categorization. We compare the categorization performance of SVM classifier based on original text feature without query expansion (*SVM_T*), majority voting by related videos (*Relevance*), and majority voting by user videos (*Interest*). The SVM classifiers are trained based on text features on the training set, and then predict the testing data. LIBSVM [6] is used to train the classifiers and RBF kernel is adopted. Each classifier outputs a confidence score of whether the input video belonging to the given category according to the provided method.

The overall performance comparison on two test sets is listed in Table 2, and a detailed performance comparison is illustrated in Figure 4. Since terms from title and tags are noisy and ambiguous for web applications, the overall performance of SVM_T is poor. However, for certain categories, such as “Gaming”, “Autos”, text is still a useful resource for classification. The user-supplied text words can provide direct clues for the video category type. If the title and tags are ambiguous and noisy, the text classifiers become incompetent. Fortunately, the social web is accompanied with rich contextual information. We can resort to related videos and user videos.

Generally, the data-driven majority voting achieves pretty good results. As shown in Figure 4 and Table 2, majority voting generally demonstrates better performance than text classification. And for some categories, the precision of data-driven methods is over 0.8. The average precision for related videos and user videos are 0.483 and 0.597 in TS_A, while 0.629 and 0.628 in TS_B, respectively. Although the idea of majority voting from related videos and user videos is simple, they give useful indication for the video categories from a global view. Beneficial from large number of “voted” videos, the biases or errors caused by a single video can be minimized. For instance, the videos uploaded by fans of Pittsburgh Steelers are mostly related to football, Steelers or NFL. At the same time, the similar information can be referred through related videos. The dominant category indicates the preference of video relevance or user interest. Another example is shown in Figure 3. The majority voting of relevance and interest indicates that most of the related videos and user videos are about Music, so the given video is highly possible to be “Music” category.

Table 2 Overall performance comparison for web video categorization.

Datasets	SVM_T	Relevance (R)	Interest (I)	QE_R	QE_U	QE_RU	SVM_T + R + I	QE_R + R + I	QE_U + R + I	QE_RU + R + I
TS_A	0.371	0.483	0.597	0.438	0.427	0.475	0.638	0.645	0.646	0.648
TS_B	0.475	0.629	0.628	0.589	0.511	0.623	0.749	0.751	0.747	0.748

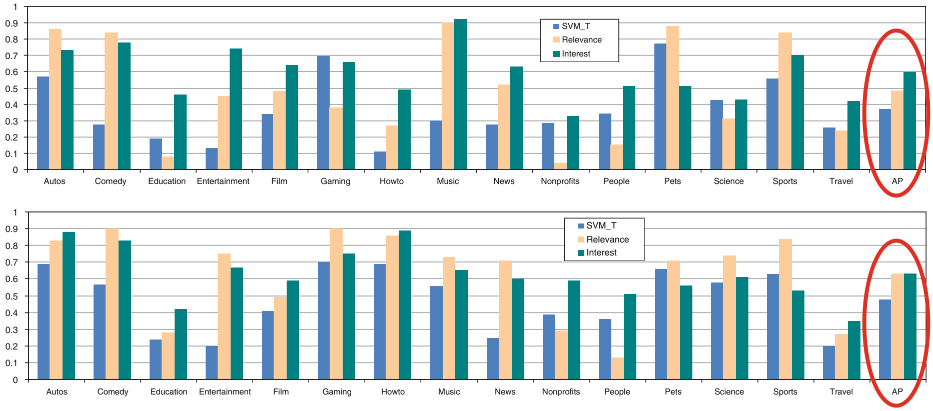


Figure 4 Performance comparison for model-based and data-driven approaches on TS_A (upper part) and TS_B (lower part) datasets.

4.4 Boosting by query expansion

In this section, we investigate the text classification improvement by query expansion from related videos and user videos. We compare the performance of SVM classification on original text feature (*SVM_T*), and SVM classification with query expansion, i.e., reinforced by related videos (*QE_R*), user videos (*QE_U*), and their combination (*QE_RU*). The classifier on original text feature (SVM_T) acts as the baseline. The performance comparison of query expansion is listed is Table 2 and shown in Figure 5.

It is understandable that the classification performance based on original noisy text information is poor. Figure 3 shows an example of a video on guitar lesson. The classifiers on original text falsely classify this music-oriented video as "Travel" category (see Figure 3 (a)), since title and tags contain terms like "Hotel", "California". To enhance the robustness of text classification, the idea of query expansion is integrated into our framework. From Table 2, we can see that significant improvement has been achieved by query expansion.

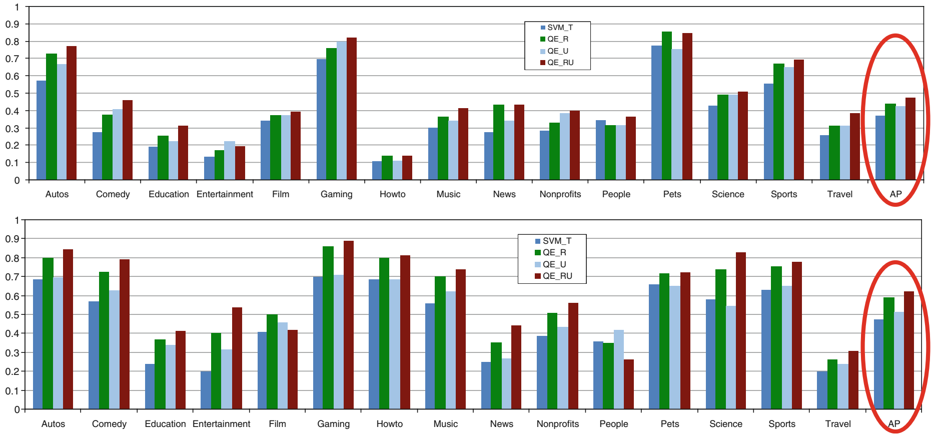


Figure 5 Performance comparison for query expansion on TS_A (upper part) and TS_B (lower part) datasets.

The classification performance has been boosted around 30%. With the assistance of query expansion, additional terms are extracted from related/user videos. It derives a better formulation of term vectors, by appending additional keywords to the initial text vector and adjusting the original weights. In our strategy, extracted terms not only append new terms to the original vector, but also adjust the weights of existing terms, so that the new formed vector is more discriminable for different categories. Table 3 demonstrates two examples with the original text terms and the top 10 expanded terms through query expansion. Query expansion can overcome the problem of lacking sufficient text terms. Useful terms are extracted from related/user videos and then update the original vectors. As seen from Table 3, in addition to existing terms, such as “iphone”, “apple”, and “3g”, additional terms, such as “macbook”, “nokia”, and “touch” are induced from query expansion, from which we can see that the extracted terms are useful and meaningful. The query expansion not only appends new terms, but also reweights the existing terms. The expanded text vector results in a higher possibility to classify the video as “Science and Technology”. The scheme of query expansion effectively compensates the shortcoming of short and noisy user-supplied text information, producing a very strong increase in the categorization quality. The contribution of query expansion is significant.

We also investigate the effect of title and tags separately for query expansion. Experiments show that tags are more useful than title, and their combination achieves the best performance. Due to space limitation, the details are ignored in this paper. Unless otherwise stated, the results of query expansion denote the combination of title and tags.

4.5 Boosting by fusion

Furthermore, semantic meaning, video relevance, and user interest, provide category confidence from different viewpoints. Their combination can give a more robust recommendation. *Semantics* comes from the model-based text classification, while *Relevance* and *Interest* are derived from the majority voting of related videos and user videos. Table 2 gives the performance of fusing Semantics, Relevance and Interest. Their combination achieves better performance than the individual ones. For example, the video showed in Figure 3 is falsely classified as “Travel” according to text information. Nevertheless, the majority voting indicates that video relevance and user “swanner” have strong preference on music. Combining the recommendation from these three aspects, the video is correctly suggested as “Music” category. Model-based classification considers the semantic meaning of text features at a relatively low level, while data-driven majority voting considers the general preference or trend at a higher level. They can complement

Table 3 Examples of original terms and the terms after query expansion (* numbers denote weights).

Videos	Video Info	Original Text / Terms from Query Expansion
	Title: iPhone 3GS vs. iPhone 3G	Original: iphone gs 3g apple camera browser performance compass speed
	Category: Science and Technology	Query Expansion: iphone:0.46 apple:0.38 3gs:0.38 3g:0.37 unboxing:0.34 macbook:0.23 nokia:0.21 tablet:0.20 touch:0.19 dell:0.18
	Title: Drake – Best I Ever Had (OFFICIAL)	Original: drake official
	Category: Music	Query Expansion: jackson:0.45 wayne:0.43 lil:0.43 michael:0.41 feat:0.35 fabolous:0.33 rick:0.33 ross:0.33 drake:0.33 boom:0.25

with each other. Their combination further boosts the performance, where its average precision reaches 0.65 and 0.75 for TS_A and TS_B datasets, respectively. The improvement is as high as 60% compared to the classification based on text features. Nevertheless, the difference among different strategies becomes minor. The contribution by query expansion from related/user videos has been embodied in the performance improvement after fusion of semantics, video relevance and user interest. The general trend after fusing Semantics, Relevance and Interest is consistent with the majority voting of Relevance and Interest. Therefore, the overall performance after fusing among different approaches is negligible.

Generally, fusion of three components improves the performance. However, this is not always true. There is the possibility that the performance using only text features is better than fusion of text and related/user videos. When a user with preference on a certain category uploads a video with different category label, the recommendation according to text classification could be inconsistent with the user interest. In this case, fusion of semantics, video relevance and user interest could have poorer performance than using only text features. For example, user "redpepper5031" uploads a video belonging to "Autos" category. Based on title and tags, text classifier assigns the correct category label "Autos" to this video. Nevertheless, the videos for user "redpepper5031" are dominated by "Entertainment". Based on the distribution of user videos, category "Entertainment" has a high score. After fusion of model-based and data-driven results, the category "Auto" is overwhelmed by the major category "Entertainment", which leads to a wrong category label.

5 Conclusion

Web video categorization is a challenging issue due to its high diversity in terms of subject, quality, and style in web scenario. User-supplied title and tags are usually too short and ambiguous to accurately describe the major content of videos, therefore, leading to a poor classification performance. The social web provides a platform with rich contextual resources supplement to the noisy text features. In this paper, related videos and user videos are utilized to improve the effectiveness of web video classification. Query expansion is adopted to boost the performance by expanding terms collected from related/user videos. The model-based classification is then built based on the new generated vectors. On the other hand, the data-driven approaches become feasible with the abundant amounts of social media and their associated rich metadata. Data-driven majority voting of category labels for related videos and user videos gives preference at a higher level, which provides constructive indication of video categories. Furthermore, the integration of semantics, video relevance, and user interests further improves the performance. Experiments on large-scale YouTube videos demonstrate the effectiveness of the proposed approach. A few useful lessons have been learnt from the experiments.

- Although user-supplied text information is short, ambiguous and noisy, it is still a useful feature for web video classification. For certain categories, it achieves good performance.
- Related videos and user videos are useful recourses for web video categorization.
- Query expansion induced from related/user videos significantly boost the performance of model-based text classification.
- Data-driven is a feasible scheme for web applications. Majority voting of related/user videos provides meaningful hints for video category.

- The information from semantics, video relevance, and user interests complements each other. Their integration further improves the performance.
- The contextual features are easy to acquire, easy to use, and scalable. Therefore, the proposed solution has high potential to be applicable to the web-scale video categorization.

There are several directions to pursuit for our future work. In this work, we assume that all related videos and user videos are relevant to the given video. This situation may not be always true. One user may have multiple interests, and the videos uploaded by the same user may cover various categories. In the future, the negative relevance feedback will be considered to reduce the impact of irrelevant videos. In this work, we deploy simple combination of original vector with query expansion, and linear fusion of model-based and data-driven approaches. We do not deliberately tune the fusion parameters. The scheme of parameter selection will be under our consideration. Furthermore, we will investigate the pure visual content-based classification in the situation that text features are totally absent.

Acknowledgement The work described in this article was supported by the National Natural Science Foundation of China (NSFC grant no. 61071184, 60972111, 6103600); Ph.D. Programs Foundation of Ministry of Education of China (Grant no: 20100184120009); the Open Project Program of the National Laboratory of Pattern Recognition (NLPR); the Fundamental Research Funds for the Central Universities (Project no. SWJTU09CX032, SWJTU10CX08); and a grant from the Research Grants Council of Hong Kong Special Administrative Regions, China, (CityU 119709).

References

1. Billerbeck, B., Scholer, F., Williams, H. E. E., Zobel, J. Query Expansion using Associated Queries. *ACM CIKM*, New Orleans, USA, pp. 2–9 (2003)
2. Brezeale, D., Cook, D.J.: Automatic video classification: A survey of the literature. IEEE Trans. Syst. Man Cybern. **38**(3), 416–430 (2008)
3. Broder, A., Ciccolo, P., Gabrilovich, E., et al.: Online Expansion of Rare Queries for Sponsored Search, pp. 511–518. WWW, Madrid (2009)
4. Cao, G., Nie, J.-Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback, pp. 243–250. ACM SIGIR, Singapore (2008)
5. Cao, J., Zhang, Y.-D., Song, Y.-C. et al. MCG-WEBV: A Benchmark Dataset for Web Video Analysis. *Technical Report*, ICT-MCG-09-001, 2009. Available at: http://mcg.ict.ac.cn/chengguo1.html
6. Chang, C.-C., Lin, C.-J.: *LIBSVM: a library for support vector machines*, 2001. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. Chrita, P.-A., Firan, C.S., Nejdl, W.: Personalized query expansion for the web. *ACM SIGIR* (2007)
8. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Trans TKDE **19**(3), 370–383 (2007)
9. Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y.: Query Expansion by Mining User Logs. IEEE Trans TKDE **15**(4), 829–839 (2003)
10. Jiang, Y.-G., Ngo, C.-W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. *ACM CIVR*, pp. 494-501 (2007)
11. Lavrenko, V., Croft, W.B.: Relevance-based language models. *ACM SIGIR*, pp. 120-127 (2001)
12. Lowe, D.: Distinctive image features from scale-invariant key points. IJCV **60**, 91–110 (2004)
13. Moxley, E., Mei, T., Manjunath, B.S.: Video annotation through search and graph reinforcement mining. *IEEE. Trans. on Multimedia* (2009)
14. Natsev, A., Haubold, A., Tesic, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. *ACM MM*, pp. 991–1000 (2007)
15. Qi, G.-J., Hua, X.-S., et al.: Correlative multi-label video annotation, pp. 17–26. ACM MM, Germany (2007)
16. Rahman, M.M., Bhattacharya, P.: Image retrieval with automatic query expansion based on local analysis in a semantical concept feature space. CIVR, Greece (2009)

17. Ramachandran, C., Malik, R., Jin, X., Gao, J.: VideoMule: A consensus learning approach to multi-label classification from noisy user-generated videos, pp. 721–724. ACM MM, Beijing (2009)
18. Rocchio, J.: Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323 (1971)
19. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans CSVT **8**(5), 644–655 (1998)
20. Shokouhi, M., Azzopardi, L., Thomas, P.: Effective query expansion for federated search, pp. 427–434. ACM SIGIR, Boston (2009)
21. Siersdorfer, S., Pedro, J.S., Sanderson, M.: Automatic video tagging using content redundancy. *ACM SIGIR*'09, pp. 395–402.
22. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Trans PAMI **30**(11), 1958–1970 (2008)
23. Wang, X., Fang, H., Zhai, C. A Study of Methods for Negative Relevance Feedback. *SIGIR'08*, pp. 219–226.
24. Wang, X.-J., Zhang, L., Li, X., Ma, W.-Y.: Annotating Images by Mining Image Search Results. IEEE Trans PAMI **30**(11), 1919–1932 (2008)
25. Xu, J., Croft, B.: Query Expansion Using Local and Global Document Analysis. *ACM SIGIR*, 2006, pp. 4–11.
26. Yan, R., Hauptmann, A.G., Jin, R. Negative Pseudo-relevance Feedback in Content-based Video Retrieval. *ACM MM*, pp. 343–346, (2003)
27. Yang, L., Liu, J., Yang, X., Hua, X.-S. Multi-Modality Web Video Categorization. *MIR'07*, pp. 265-274.
28. Yuan, X., Lai, W., Mei, T., Hua, X.-S., Wu, X.-Q.: Automatic Video Genre Categorization using Hierarchical SVM. ICIP, Atlanta (2006)
29. Zhai, C., Lafferty, J.D. Model-based Feedback in The Language Modeling Approach to Information Retrieval. *ACM CIKM*, pp. 403–410, (2001)