DOI:10.11992/tis.201603040
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0957.032.html

一种多模态融合的网络视频相关性度量方法

温有福[1,2],贾彩燕[1],陈智能[2]
(1.北京交通大学 交通数据分析与数据挖掘北京市重点实验室,北京 100044; 2. 中国科学院自动化研究所 数字内容技术与服务研究中心,北京 100190)

摘 要:随着网络和多媒体技术的发展,视频分享网站中的网络视频数量呈爆炸式增长。海量视频库中的高精度视频检索、分类、标注等任务成为亟待解决的研究问题。视频间的相关性度量是这些问题所面临的一个共性基础技术。本文从视频视觉内容,视频标题和标签文本,以及视频上传时间、类别、作者 3 种人与视频交互产生的社会特征等多源异构信息出发,提出一种新颖的多模态融合的网络视频相关性度量方法,并将所获相关性应用到大规模视频检索任务中。YouTube 数据上的实验结果显示:相对于传统单一文本特征、单一视觉特征的检索方案,以及文本和视觉特征相融合的检索方案,文本视觉和用户社会特征多模态融合方法表现出更好的性能。

关键词:网络视频; 海量视频;社会特征; 交互;多源异构信息;多模态信息融合; 相关性度量; 视频检索
中图分类号:TP393 **文献标志码**:A **文章编号**:1673-4785(2016)03-0359-07

中文引用格式:温有福,贾彩燕,陈智能.一种多模态融合的网络视频相关性度量方法[J]. 智能系统学报, 2016, 11(3): 359-365.
英文引用格式:WEN Youfu, JIA Caiyan, CHEN Zhineng. A multi-modal fusion approach for measuring web video relatedness[J]. CAAI transactions on intelligent systems, 2016,11(3): 359-365.

A multi-modal fusion approach for measuring web video relatedness

WEN Youfu[1,2], JIA Caiyan[1], CHEN Zhineng[2]
(1. Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China; 2. Interactive Media Research and Services Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: With the advances in internet and multimedia technologies, the number of web videos on social video platforms rapidly grows. Therefore, tasks such as large-scale video retrieval, classification, and annotation become issues that need to be urgently addressed. Web video relatedness serves as a basic and common infrastructure for these issues. This paper investigates the measurement of web video relatedness from a multi-modal fusion perspective. It proposes to measure web video relatedness based on multi-source heterogeneous information. The multi-modal fusion simultaneously leverages videos´ visual content, title, and tag text as well as social features contributed by human-video interactions (i.e., the upload time, channel, and author of a video). Consequently, a novel multi-modal fusion approach is proposed for computing web video relatedness, which serves to give a ranking criterion and is applied to the task of large-scale video retrieval. Experimental results using YouTube videos show that the proposed text, visual, and users´ social feature multi-modal fusion approach performs best in comparison tests with three alternate approaches; i.e., those approaches that compute web video relatedness based just on text features, just on visual features, or jointly on text and visual features.

Keywords: web video; large-scale video; social feature; human-video interactions; multi-source heterogeneous information; social features; multi-modal fusion; relatedness measurement; video retrieval

视频是集图像、声音和文字信息于一体的多源信息载体,其丰富直观的表达形式非常契合人类接受信息的方式。随着网络和多媒体技术的快速发展,在线视频服务正在以不可阻挡之势在互联网平台上蓬勃发展。成立于 2005 年的视频分享网站 YouTube,目前已成为世界第三大网站和第二大搜

收稿日期:2016-03-19. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金项目(61473030,61303175);重点大学研究基金项目(2014JBM031);重点实验室数字媒体技术开放课题
通信作者:贾彩燕. E-mail:cyjia@bjtu.edu.cn.

索引擎。在国内,主流视频分享网站优酷网目前拥有超过 1 亿网络视频,日均观看次数超过 4 亿次。网络视频已成为社会生活中知识传播、信息获取和休闲娱乐的重要载体之一。

网络视频数量的持续快速增长使得海量网络视频库中视频相关性的快速准确度量成为一个至关重要的课题。对视频分享网站而言,若能更好更快地度量网络视频之间的相关性,视频推荐、视频检索、视频主题发现等典型视频服务则可以得到更好的开展。对视频网站上网络视频的检索而言,最为核心的是相关性度量问题,即对给定的查询视频,挖掘它与库中其他视频在文本或视觉上的相似性,进而得到相关度打分,通过相关度分数的高低来得到检索结果。关于相关性度量,有使用视频文本相似度的方法,如 Zhu 等[1]应用关键词投票的相似性度量方法进行视频文本标题的分类。有文本相似性和视频相似性相结合的方法,如 Brezeale 等[2]结合文本相似度和视觉特征相似度进行的视频匹对与分类;Schmiedeke 等[3]融合视频的文本标签和视觉相似度进行视频的分类。但主流研究集中在基于视频内容的相似性计算上,包括各类特征的提取、检索结果的求精和加速等[4-9]。

近年来,人们开始引入更多的模态和信息来加强视频相似性度量的准确性。Feng 等[10]融合视频标注、视觉、视频间关系来提升视频检索质量;Brezeale 等[11]面向视频分类,分析比较了文本、视频、音频三模态融合的方法。上述方法虽然在实践中取得了很好的效果,但忽视了网络视频网页上的各种信息,例如视频类别、视频上传时间、视频作者等,我们将这些特征称为社会特征。显然,这些信息从特定角度体现了视频内容,可以用来更好地度量视频之间的相关性。已有研究表明:仅仅使用视频的视听觉内容很难将视频归到某一种类[12]。而网络视频网页周边的相关信息提供了很多的资源,这些资源可以更加准确地评判视频之间的相似性,从而有利于检索实现。目前也已经有学者开始利用这些信息去研究网络视频的分类和检索问题。例如 Wu 等[13]结合用户兴趣和文本解决视频分类问题;Davidson 等[14]提出利用视频的“共同观看”关系进行视频推荐。上述成果表明,网络视频网页上蕴含了丰富的信息,可以被用来更好地计算两个视频之间的相关度,从而为更多模态的融合提供更好的思路。上述多模态的融合方法,或者增加音频、或者增加文本、或者使用视频共现关系,但是没有全面考虑社会特征对相关度度量的影响。

本文将文本(视频标题、视频标签)、视觉(视频内容)、社会特征(视频的上传时间、视频的作者、视频的类别)进行全面的多模态的融合。真实网络视频数据上的实验表明:相比于仅考虑视觉、文本、社会或者是视觉+文本的方法,本文方法可以取得 5%~25%的性能提升,充分说明了本文方法的有效性。

1 多模态融合的网络视频相关性度量

图 1 给出了本文多模态融合的网络视频相关性度量方法的整体框架。

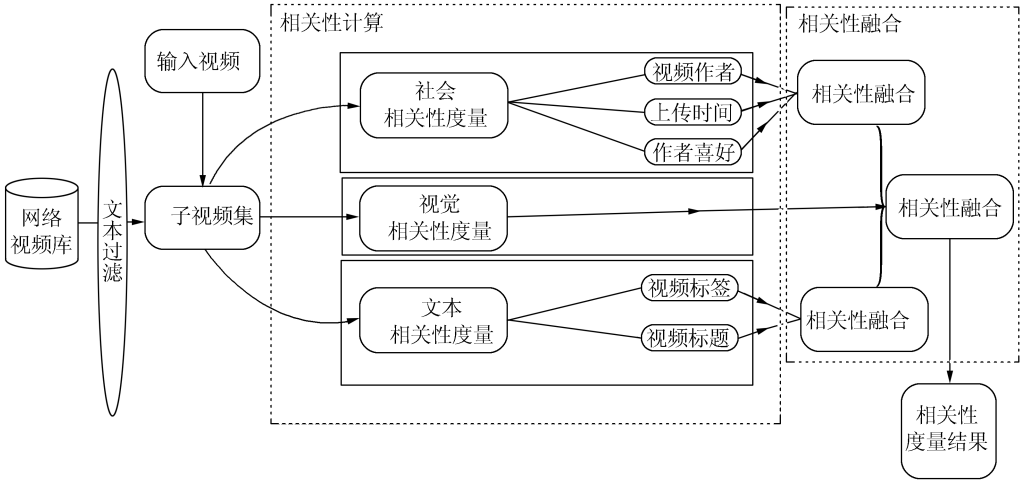

图 1 检索融合流程图

Fig.1 The flow chart of image retrieval

多模态融合网络视频相关性度量方法分为 3 个模块:1)文本过滤;2)相关性度量;3)相关性融合。

对给定的网络视频,首先利用文本过滤模块,滤除掉大量不相关的网络视频。然后,分别通过社会相关性度量、文本相关性度量和视觉相关性度量 3 个子模块,计算 3 个模态上的相关度。最后,通过相关性融合模块,进行融合,从而实现更准确相关性度量。

1.1　基于文本的过滤

真实网络视频库都拥有大量视频,且对任意给定的输入视频,库中通常仅有一小部分视频与其相关。为避免大量相关性计算耗费在无意义的视频上,提出如下所述的一种基于文本的过滤策略。令 $T_q=\{t_1^q,t_2^q,\cdots,t_k^q\}$ 为输入视频 v_q 所在网页上标题和标签中的关键字集合,则整个视频库 Ω 中所有满足式(1)的视频都将被滤除。

$$\{v_i \in \Omega \mid T_q \cap T_i = \varnothing\} \tag{1}$$

上述步骤假设:两个网络视频的周边文本中应有至少一个共同的关键字才值得进行相关度计算。实际情况中,这一假设对绝大部分的相关视频而言都是成立的。通过这一操作,大量无关的网络视频得到了有效滤除。

令文本过滤后生成的视频集合为 Ω^q,Ω^q 中视频虽然与输入视频有至少一个相同的关键字,但其中仍有大量无关或相关性不高的视频。接下来,我们将从社会特征、文本和视觉 3 个模态上进行相关度挖掘和分析。首先介绍社会特征方面的度量方法。

1.2　社会特征的选择与相关性计算

社会特征是网络视频区别于其他视频的特色之一,它泛指人与视频各种交互行为产生的数据集合。典型的社会特征包括视频上传时间、上传作者、视频类别、视频观看次数、视频评论等。

从社交网络和多媒体的角度讲,社会特征是用户与视频网页进行的交互以及多媒体服务人员为相应的视频或者图像所做的对于视频或者图像内容的文字反映。社会特征包含的内容很多,例如用户上传视频的时间、视频的分类类别、视频的观看次数、视频的作者等。这些社会特征反映了用户个人的具体信息,用户与用户的关联信息等。通过用户个人内部的关联信息以及用户与用户之间的关联性社交网络,可以更容易发现相关性视频,进而更好地进行视频的相似性度量。本文选用视频的上传时间(反映视频的新颖性)、视频的分类类别、视频上传作者 3 种特征进行社会特征融合的尝试。

1.2.1　上传时间

视频的上传时间是反映用户对于当前所关注视频的时间在特定时间段上的描述。例如:存在 3 个视频 v_q、v_i 和 v_j,如果 v_q 与 v_i 和 v_j 之间的时间间隔分别为 3 个月和 3 年,那么 v_q 与 v_i 之间的相似度会更高一些。

为度量视频的这种相关性,我们首先计算视频 v_q 与 v_i 的时间差值 t_{iq}:

$$t_{iq}=\mathrm{abs}(t_{v_q}-t_{v_i})$$

然后,令时间间隔 Δt 为 0.5 个月,将 v_q 经过文本过滤后生成的视频集合 Ω^q 等间隔划分为 n 个不相交的子集合:

$$\Omega^q=\bigcup_{j=0}^{n}\Delta\Omega_j^q \tag{2}$$

$$\Delta\Omega_j^q=\{v_i \mid t_{iq} \in [j\times\Delta t,(j+1)\times\Delta t)\}$$

则 $\Delta\Omega_j^q$ 为与 v_q 时间差落在 $[j\times\Delta t,(j+1)\times\Delta t)$ 区间的视频集合,式(2)中令 $n=7$。考虑到与 v_q 的时间差越近的视频应有更高的相关度权值,将 $\Delta\Omega_j^q$ 的权值指派为 $n-j$,则视频在上传时间方面的相关度可通过式(3)计算:

$$f_{\mathrm{time}}(v_q,v_i)=\frac{(n-j)}{\sum_{j=0}^{n}j}M(\Delta\Omega_j)$$

$$M(\Delta\Omega_j)=\begin{cases}1, & v_i \in \Delta\Omega_j\\ 0, & \text{其他}\end{cases} \tag{3}$$

式(3)实质上是依照特定的时间间隔对视频进行划分,距离给定视频 v_q 上传时间越近则 f_{time} 值越高。

1.2.2　视频类别

常见的视频类别包括财经类、政治类、综艺类等。一般地,属于同一个类别的视频的相关度通常更高。例如:有 3 个视频中有两个是属于综艺类的,而另一个是属于政治类的。从视频类别的层面分析,两个综艺类视频的相似性程度应该更高一些。

对此,我们应用式(4)反映:

$$f_{\mathrm{channel}}(v_q,v_i)=\begin{cases}1, & (\mathrm{channel}_q=\mathrm{channel}_i)\\ 0, & \text{其他}\end{cases} \tag{4}$$

式中$\mathrm{channel}_q$ 是视频 v_q 所属的类别。式(4)反映了依照视频的类别对视频进行划分,落入到同一个类别的视频则其 f_{channel} 值为 1。

1.2.3　上传作者

每位用户都有自己的兴趣爱好,这一特点通常在他/她上传的视频集合上可以得到一定体现。例

如,用户 a_q 和 a_i 上传的视频主要是体育类的,而用户 a_j 上传的视频主要是财经类的,在不考虑其他因素的情况下,a_q 上传的视频与 a_i 上传的视频的相关度通常比他/她与 a_j 上传的视频的相关度更高。因此,本小节我们首先建立视频作者的视频喜好模型,然后通过喜好模型的相似度度量不同视频在上传用户这一因素上的相关性。

设视频网站有 K 个视频类别,则用户 a_q 上传的所有视频可表示为一个 K 维喜好向量:

$$\boldsymbol{A}_q = (a_q^1, a_q^2, \cdots, a_q^K)$$

式中 a_q^i 为用户 a_q 上传的所有视频中,标记为类别 i 的视频的数量。基于此,两个视频在上传用户层面上的相关度可定义为

$$f_{\text{author}}(v_q, v_i) = f_{\text{author}}(\boldsymbol{A}_q, \boldsymbol{A}_i) \tag{5}$$

式中$\boldsymbol{A}_q$、$\boldsymbol{A}_i$为归一化后的作者喜好向量。

式(5)反映了若作者 $\boldsymbol{A}_q$ 与作者 $\boldsymbol{A}_i$ 有着类似的喜好向量,则他们上传视频的相关性在上传作者这一维度应被赋予更大的值。这一思想与著名的协同推荐方法有着类似之处。

1.3.4 多社会特征融合

以上 3 个社会特征对于视频的相似性度量都起到有益效果,我们将每个社会特征做相应归一化后进行融合,公式为

$$F_{\text{social}} = \alpha f_{\text{time}} + \beta f_{\text{author}} + (1 - \alpha - \beta) f_{\text{channel}}$$

式中 $\alpha, \beta, 1-\alpha-\beta \in [0,1]$ 分别为上述 3 种社会特征的权重。本文在实验中为三者赋予同等的权重。

1.3 文本相关性计算

在文本相关性计算的过程中,我们采用了传统基于文本关键词匹配的方法。该方法首先基于整个视频库 $\boldsymbol{\Omega}$ 中所有视频的标题和标签构建一个 N 维向量空间模型。然后,基于该模型并结合文档处理中常用的 TF-IDF 加权方法,将每个视频的标题和标签关键字集合分别表示为一个 N 维标题特征向量和标签特征向量,通过计算不同视频间标题和标签特征向量的相似性,实现对它们文本相关性的评判。

1.3.1 标题

标题是对视频主要内容的高度概括,通常简要描述了视频事件的主要内容。基于相关视频在内容上也存在较强相关性的假设,两个强相关的视频标题关键词通常也会有一定交集。这种性质则可以反映为

$$f_{\text{title}}(v_q, v_i) = \cos(\boldsymbol{f}_q^{\text{title}}, \boldsymbol{f}_i^{\text{title}}) \tag{6}$$

式中 $\boldsymbol{f}_q^{\text{title}}$ 为视频 v_q 的标题特征向量。

通过式(6),如果视频 v_q 与视频 v_i 对应的标题向量 $\boldsymbol{f}_q^{\text{title}}$ 和 $\boldsymbol{f}_i^{\text{title}}$ 越相似,则它们的余弦相似度越大,亦即在标题方面的相关性 f_{title} 越大。

1.3.2 标签

相对于标题,标签通常则更加具体地反映了视频的内容,例如涉及的人物、地点和其他专有名词等。与标题类似,两个强相关视频的标签关键词集合存在一定交集的情况也很常见。因此,我们通过与标题类似的公式反映这种性质:

$$\boldsymbol{f}_{\text{tag}}(v_q, v_i) = \cos(\boldsymbol{f}_q^{\text{tag}}, \boldsymbol{f}_i^{\text{tag}})$$

式中 $\boldsymbol{f}_q^{\text{tag}}$ 为视频 v_q 的标签特征向量。同理,两个视频的标签向量 $\boldsymbol{f}_q^{\text{tag}}$ 和 $\boldsymbol{f}_i^{\text{tag}}$ 越相似,则它们在标签方面的相关性 $\boldsymbol{f}_{\text{tag}}$ 越大。

1.3.3 文本信息融合

以上标题和标签的信息从不同角度体现了视频的内容,将它们通过式(7)融合起来,可以更为准确地描述了两个视频在文本模态上的相似度:

$$F_{\text{text}} = \gamma f_{\text{title}} + (1 - \gamma) f_{\text{tag}} \tag{7}$$

式中 γ、$1-\gamma \in [0,1]$ 分别为标题和标签。在实验中 γ 被经验性的设置为 0.5。

1.4 视觉相关性计算

视频的视觉内容是用户认知的根本性来源,它从本质上反映视频的相似性程度。由此,视觉相似性计算对于视频的相关性度量是不可或缺的一个方面。在视觉的相关性度量方面,我们采用了 Zhao 等[15]提出大规模拷贝视频检测方法计算两个给定视频之间的视觉相似度。

该方法首先计算任意两个视频帧之间的视觉相似度,再根据 Hough 变换原理,对相似度大且在时序上有较高一致性的视频赋予更大的相似度,该相似度基本反映了两个视频之间的程度。在计算视频帧之间的两两相似度方面,采用的是经典基于视觉词袋模型的方法。其基本流程如图 2 所示。

图 2 视觉词典方法

Fig.2 The method of visual vocabulary

具体地,我们首先以 5 s 为时间间隔,对视频进行等间隔采样,提取相应图像帧。其次,在图像帧上提取 sift 特征,并利用预先训练好的视觉特征词典将每个图像帧的 sift 特征集合量化为一个高维的稀疏向量,进而通过比较两个向量的相似性实现对图

像帧相似度的计算。

为实现从图像帧相似度到视频相似度的计算,Zhao 等使用了 Hough 变换投票方法。其描述为:1)以等时间戳进行时间对准;2)进行等时间间隔霍夫投票,即如果两个视频某个相同时间差片段间的视觉相似性高于某一个阈值,则予以投票;3)以时间差为横轴形成的直方图,直方图越高,则两个视频连续片段越相似,亦即视频越相似。

除 Hough 变换投票外,从图像帧相似度到视频相似度计算方面的典型方法还有基于网络流约束的线性规划的方法[16]等。考虑更多方法并将它们的优势结合起来,实现更加准确的视频视觉相似度计算也是我们的下一步工作之一。

1.5　多模态相关性计算

通过上述介绍的方法,可以分别计算得到两个视频在社会特征、文本和视觉模态的相似度,通过式(8)对这 3 方面的相似度进行融合:

$$F_{\text{fusion}} = \omega F_{\text{social}} + \tau F_{\text{text}} + (1 - \omega - \tau) F_{\text{visual}} \quad (8)$$

式中 ω、τ、$1-\omega-\tau \in [0,1]$ 分别为社会、文本、视觉 3 种异构多模态特征的融合权重。在进行融合之前,我们将各个模态信息进行相应的归一化操作,每个模态信息均取相应均等权重。

以上对本文方法的各个模块进行了详细介绍。本文初步探讨多种在不同层次、从不同角度反映视频相关性的信息的融合方法。在社会特征方面融合了视频上传日期、类别和作者信息,文本方面则融合了视频标题和标签的相似性;在整体层面,则将社会、文本和视觉三大模态的异构信息进行了融合。目前,对于各模态内部以及各模态之间权重的分配方面暂时没有展开过多细节讨论。

2　实验结果及分析

2.1　实验数据准备

为验证本文方法的有效性,我们在 MCG-WEBV 2.0[17]数据集上进行了实验。该数据集按月收集了 2008 年 12 月-2009 年 11 月间 YouTube 给定的 19 个视频类别上的“每月观看最多”视频,以及它们的相关视频和同作者视频,共计 248 887 个。

上述数据收集方法使得这一年内网络空间的热点事件在该数据集上都有所体现。基于此,选择发生在这段时间内关于 11 个热点话题的视频进行实验,表 1 列出了这些话题的基本情况。

表 1　热点话题的基本情况

Table 1 The hot topics

ID	话题描述	日期
1	Bush was attacked by shoes in Iraq	200811
2	Obama's inauguration speech	200901
3	Susan Boyle in Britain's Got Talent	200904
4	Amanda Holden in Britain's Got Talent	200904
5	The death of Michael Jackson	200906
6	Cenk Uygur's news show	200908
7	Lady Gaga Poker Face	200902
8	Silvio Berlusconi was attacked in Milan	200911
9	Brad Pitt's movies	200902
10	Lauren Luke's makeup	200904
11	Miley Cyrus's show	200812

表 2 给出了实验中依照特定关键词筛选条件进行文本过滤后,每个话题剩余的满足条件的视频数量。第 1 列(ID)表示的是视频的 11 个话题的集合,表的第 2 列反映了每一个主题下的关键词,最后 1 列反映了依照筛选条件,时间间隔为依照时间(Date)参照点前后 4 个月的条件下,所得到的视频数量结果。相对于整个视频库中总的视频数量,本文提出的文本过滤策略平均滤掉了 99.82%的视频数据。从而使得视觉等较为耗费计算资源的运算可以在一个相对较小的集合上进行。

首先,我们使用类似文献[18]的方法,对表 1 中的每一个话题给定文本筛选条件,利用 1.2 节中描述的文本筛选方法筛选出与这几个话题相关的视频集合,见表 2。

表 2　特定筛选条件及筛选结果

Table 2　The results of specific conditions

ID	关键词	数量
1	bush shoe	210
2	obama inauguration	436
3	Susan Boyle	683
4	Amanda Holden	294
5	Michael Jackson	1923
6	cenk uygur	276
7	Lady Gaga poker	196
8	Silvio Berlusconi	413
9	brad pitt	150
10	lauren luke	149
11	Miley Cyrus	467

然后,对候选集合中的视频进行人工标注,一个视频被标注为与该话题相关当且仅当它包含了描述

上述热点事件的镜头。两个视频的相关度被置为 1(即标注为相关视频)当且仅当它们都被标注为与同一个话题相关,其余情况下,两个视频的相关度被置为 0,即不相关。对任意一个输入视频,依据 2.2.2 小节给出的不同方法,计算它与数据集中其他视频的相关度,为每个方法得到一个按相关度得分从高到低排序的结果列表。然后,我们用多媒体检索中广泛采用的 AP 作为衡量结果相关性度量准确性的指标,其中 AP 的计算公式为

$$\mathrm{AP} = \frac{1}{n^+}\sum_{j=1}^{n} I_j \times \frac{R_j}{j}$$

其中 n^+是测试集中相关视频的总数,$n = 100$ 表示仅考虑列表的前 100 个结果。若第 j 个视频是相关视频,则 $I_j = 1$,否则 $I_j = 0$。R_j 表示前 j 个结果中相关视频的数量。实验中,所有被标注为与某个话题相关的视频组成了输入视频集合。它们被一一作为输入视频,通过 2.2.2 小节的各种方法得到相应的相关视频结果列表并计算 AP。为便于结果展示和分析,我们将每个话题所有相关视频的 AP 进行平均,得到该话题的 AP。然后,再对多个话题 AP 再次求平均,得到刻画各个方法整体性能的 MAP 值。

2.2 多模态融合相关性度量实验

依据第 1 节给出计算公式,可以得到任意两个视频的相关度。我们将计算得到的相关度应用到对上述话题的网络视频检索上,依据相关度大小从高到低对视频进行排序。表 3 给出了用本文多模态融合的网络视频相关度度量方法的检索结果。作为比较,仅用视觉相似性、文本相似性以及视觉与文本相结合相似性进行检索的实验结果也在表 3 中给出。

表 3 各类方法试验结果对照表

Table 3 The results of all kinds of methods

ID	Visual	Text	Visual+Text	Fusion
1	0.814	0.742	0.813	**0.831**
2	0.305	0.180	0.312	**0.330**
3	0.406	0.495	**0.410**	0.396
4	0.654	0.840	0.702	**0.740**
5	0.304	0.204	0.400	**0.503**
6	0.973	0.905	0.970	**0.972**
7	0.110	0.030	0.210	**0.230**
8	0.603	0.306	0.650	**0.704**
9	0.400	0.200	0.440	**0.514**
10	0.021	0.053	**0.301**	0.043
11	0.598	0.418	0.605	**0.714**
MAP	0.471	0.397	0.528	**0.543**

表 3 中的第 1 列 ID 表示 11 个相关话题,后面的每 1 列值表征的是 AP 的检索结果;最后 1 列反映了实验结果。我们的方法将特有的社会特征与文本、视觉进行融合,相比于单一的文本视觉,以及文本和视觉相融合的方法取得到了较好的效果。

综上所述,将视觉、文本和社会特征进行多模态融合的方案取得了上述最好的结果,相对单一视觉,单一文本方法 5%~25%的性能上的提升,相对于文本和视觉融合的方法,我们的方法也取得了更好的结果。

3 结束语

本文提出一种新颖的网络视频相似性度量方法。从文本、视觉和社会特征 3 个角度同时挖掘视频的相关关系并进行融合。在社会特征方面,我们选择了视频的上传时间、作者、类别信息,给出了相关性在这 3 种特征上的形式化度量方法;文本特征方面,在向量空间模型中分别计算了两个视频在标题和标签上的相似度;视觉特征方面,采用主流基于视觉拷贝视频检测的方法度量两个视频的相似性。通过在宏观上对上述相似度进一步融合,实现了对视频相关性的准确鲁棒度量。真实 YouTube 数据上的视频检索实验表明,相比于仅考虑视觉、文本、或是视觉和文本相结合的方法,本文方法可取得 5%~25%的性能提升。

以上工作初步证明了在相关性度量方面融合多模态信息的合理性和有效性。我们的下一步工作将在以下 3 个方面进行。1)研究更加有效的视频视觉相似性度量方法。在度量细粒度的单纯视频视觉相似性的研究,已经有两种主流方法[18],基于这些方法的改进和融合也是一个研究点;2)探索更加有效的多模态信息融合方法。融入更多社会特征,研究多模态特征融合权重的自适应确定方法;3)将本文方法计算得到的相关度应用到除检索外的更多应用场景上,例如网络视频分类、标注等。

参考文献:

[1] ZHU Weiyu, TOKLU C, LIOU S P. Automatic news video segmentation and categorization based on closed-captioned text[C]//Proceedings of IEEE International Conference on Multimedia and Expo. Tokyo, Japan, 2001: 829-832.

[2] BREZEALE D, COOK D J. Using closed captions and visual features to classify movies by genre[C]//Poster Session of the Seventh International Workshop on Multimedia Data Mining. Philadelphia, Pennsylvania, USA, 2006.

[3]SCHMIEDEKE S, KELM P, SIKORA T. TUB @ MediaEval 2011 genre tagging task: prediction using bag-of-(visual)-words approaches[C]//Working Notes Proceedings of the MediaEval 2011 Workshop. Pisa, Italy, 2011: 1-2.

[4]LAW-TO J, CHEN Li, JOLY A, et al. Video copy detection: a comparative study[C]//Proceedings of the 6th ACM International Conference on Image and Video Retrieval. New York, NY, USA, 2007: 371-378.

[5]WU Xiao, HAUPTMANN A G, NGO C W. Practical elimination of near-duplicates from web video search[C]//Proceedings of the 15th ACM International Conference on Multimedia. New York, NY, USA, 2007: 218-227.

[6]SONG Jingkuan, YANG Yi, HUANG Zi, et al. Multiple feature hashing for real-time large scale near-duplicate video retrieval[C]//Proceedings of the 19th ACM International Conference on Multimedia. New York, NY, USA, 2011: 423-432.

[7]PERRONNIN F, DANCE C. Fisher kernels on visual vocabularies for image categorization[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA, 2007: 1-8.

[8]JéGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA, 2010: 3304-3311.

[9]TAN H K, NGO C W, HONG R, et al. Scalable detection of partial near-duplicate videos by visual-temporal consistency[C]//Proceedings of the 17th ACM International Conference on Multimedia. New York, NY, USA, 2009: 145-154.

[10]FENG Bailan, CAO Juan, CHEN Zhineng, et al. Multimodal query expansion for web video search[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA, 2010: 721-722.

[11]BREZEALE D, COOK D J. Automatic video classification: a survey of the literature[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2008, 38(3): 416-430.

[12]YANG Linjun, LIU Jiemin, YANG Xiaokang, et al. Multimodality web video categorization[C]//Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval. New York, NY, USA, 2007: 265-274.

[13]WU Xiao, ZHAO Wanlei, NGO C W. Towards google challenge: combining contextual and social information for web video categorization[C]//Proceedings of the 17th ACM International Conference on Multimedia. New York, NY, USA, 2009: 1109-1110.

[14]DAVIDSON J, LIEBALD B, LIU J, et al. The YouTube video recommendation system[C]//Proceedings of the 4th ACM Conference on Recommender Systems. New York, NY, USA, 2010: 293-296.

[15]ZHAO Wanlei, WU Xiao, NGO C W. On the annotation of web videos by efficient near-duplicate search[J]. IEEE Transactions on Multimedia, 2010, 12(5): 448-461.

[16]TAN H K, NGO C W, CHUA T S. Efficient mining of multiple partial near-duplicate alignments by temporal network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(11): 1486-1498.

[17]CAO J, ZHANG Y D, SONG Y D, et al. MCG-WEBV: a benchmark dataset for web video analysis[R]. Technical Report, Beijing, China: Institute of Computing Technology, 2009: 324-334.

[18]JIANG Yugang, JIANG Yudong, WANG Jiajun. VCDB: a large-scale database for partial copy detection in videos[M]//FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision-ECCV 2014. Zurich, Switzerland: Springer, 2014: 357-371.

作者简介:

温有福,男,1991 年生,硕士研究生,主要研究方向为视频/图像检索、社交网络分析。

贾彩燕,女,1976 年生,副教授,博士生导师,博士,主要研究方向为数据挖掘、社会计算、文本挖掘及生物信息学。近年来主持国家自然科学基金面上项目 1 项,主持国家自然科学基金青年基金项目和面上项目 1 项;参加国家自然科学基金重点项目、国家科技重大专项、北京市自然科学基金项目各 1 项;获得湖南省科学技术进步二等奖 1 项,发表学术论文 40 余篇。

陈智能,男,1982 年生,副研究员,博士,主要研究方向为多媒体内容分析与检索、机器学习、图像处理。近年来主持国家自然科学基金青年基金 1 项,发表学术论文 20 余篇。