# 跨社交媒体网络大数据下的用户建模

项连城[1,2]，桑基韬[1,2]，徐常胜[1,2]
1. 中国科学院自动化研究所，北京 100190；2. 中国科学院大学，北京 100049

**摘要**
社交媒体大数据中的多源性体现在不同社交媒体网络产生的内容上，从多源的角度分析跨社交媒体网络可以将独立数据的价值通过整合其他来源和模态的数据充分挖掘和释放出来，提高大数据的利用效率。跨社交媒体网络的用户建模是分析和应用多源社交媒体大数据的重要体现。跨社交媒体网络中的多源数据共享独立用户空间，提出以用户为桥梁对多源数据进行关联挖掘，将挖掘得到的关联模式分别应用于跨社交媒体网络的用户人口属性建模和兴趣建模中，并应用到社交媒体应用的个性化服务中。

**关键词**
跨社交媒体网络；用户建模；人口属性；兴趣属性

中图分类号：TP37 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016052

# *Cross-OSN user modeling in big data*

XIANG Liancheng[1,2], SANG Jitao[1,2], XU Changsheng[1,2]
1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

***Abstract***
Social media variety mainly concerns with the contents created and consumed in different online social network (OSN). Analyzing cross-OSN from the perspective of "variety" is beneficial to exerting the potential of big data, by integrally analyzing and exploiting the multi-sourced and multi-modal data. The problem of exploiting the cross-OSN data for comprehensive user modeling, which is fundamental in the context of multi-sourced social media big data was addressed. Inspired by the fact that the cross-OSN data shares unique user space, take the users as a bridge for associations mining between OSN was proposed. The discovered association patterns were then utilized in cross-OSN user demographic attribute inference and interest modeling in cross-OSN respectively, which can be further applied to personalized social media services.

***Key words***
cross-OSN, user modeling, demographic attribute, interest attribute

## 1 引言

计算机、手机、相机等电子产品的问世和广泛普及，不但使数据有了新的载体并扩大了其内涵，而且使人们可以更方便地产生数据、传播数据，极大地加快了数据的增长速度。2004年，以Facebook为代表的社交媒体网络为人们提供了平台，供人们分享日常生活工作中的所感、所想并进行交流。这些社交媒体网络促使人们完成了从被动接受数据到主动产生并分享数据的一个重大革命，也带来了真正的数据爆炸。据统计，以社交媒体行为数据为主的非结构化数据已经占到了人类数据总量的75%，达到600 EB，其中平均每个人贡献的数据超过100 GB。人们已经进入社交媒体大数据时代，但是相比大数据的产生速度和获取能力，大数据的价值提炼和挖掘能力仍然比较低，出现了“大数据，小价值”的失衡状况。

社交媒体表现出典型的大数据“4V”特征，即volume（体量巨大）、velocity（增长迅速）、variety（异构多源）、veracity（价值密度低）。在这4个特征中，volume关注数据存取和索引的速度，velocity关注数据计算的效率，variety关注数据分析的复杂性，veracity关注数据质量。社交媒体大数据的分析和应用，需要重点研究variety这一维特征：即处理和利用不同来源、不同类型的用户产生内容（user-generated content）。variety有异构和多源两种解释：异构是指不同形式、不同类型的数据，比如文本、图像、语音、视频等传统多模态数据以及随着社交媒体出现的图片微博、语音图片等新媒体数据；多源是指不同来源的数据，比如互联网数据可能从桌面或者移动编辑采集而来，可能由官方或者个人发布上传，并存在于新闻、博客、播客、论坛等不同的网站上。异构数据的分析和应用在“小数据”时代已经得到了充分的关注，在多媒体传输、存储、特征学习、语义理解等方面取得了显著成果。与之相比，多源数据的分析和应用研究却刚刚起步。实际上，异构性和多源性都是挖掘社交媒体大数据价值的关键。跨社交媒体网络同时具有这两种性质，不同的社交媒体网络是不同的数据来源，而每个社交媒体网络都拥有大量的异构数据。分析跨社交媒体网络可以将独立数据的价值通过混合其他来源和模态的数据充分挖掘和释放出来，更加有效地提炼大数据的价值。

近年来，跨社交媒体网络分析的优势逐渐被发现和重视，各种社交媒体网络之间的合作相继展开。同种类型的社交媒体网络进行横向合作可以最大化垄断利润。仅2015年，国内发生了滴滴打车和快的打车合并、58同城和赶集网合并、美团和大众点评网合并以及去哪儿网和携程网合并，涵盖了电召出行、生活服务平台、O2O、在线旅游等各个方面。而不同类型的社交媒体网络进行纵向合作的目的是增强生态的上下游。如Google（谷歌）收购YouTube与国内腾讯收购大众点评网、百度收购糯米一样，后者为前者的下游，前者向后者导入流量。此外，Amazon（亚马逊）和Twitter也有合作，两者关系类似于淘宝网和微博，后者为前者的上游，前者通过广告等向后者导入流量。这些横向或纵向的合作，目前仍停留在浅层，即分享不同社交媒体网络的用户集合，并没有深入综合利用各网络中独立、分散的数据。其实，社交媒体网络中的多源异构数据共享独立用户空间，以用户为桥梁进行连接。以用户为中心分析跨社交媒体网络可以有效地连接各社交媒体网络中的独立异构数据，并予以综合利用，充分释放其潜在价值，从而实现各

网络的深层合作。

跨社交媒体网络的用户建模是分析和应用社交媒体大数据的重要体现。社交媒体应用的核心是信息服务，在信息内容和用户数量都爆炸式增长的今天，通过用户建模进行个性化信息服务，是高效地对接用户和内容、解决信息过载的有效途径。由于不同的社交网络应用关注不同类型的服务，同一个用户会同时参与到不同社交网络中。Global Web Index 2015的统计发现，在调查的50个社交媒体网站中，每个人平均拥有5.54个账号，并定期活跃在2.82个网络上。参与多个社交网络的用户的网络足迹是其在不同社交网络平台行为数据的聚合，彼此关联共同反映用户的属性和兴趣。进行跨社交媒体网络的用户建模就是整合用户分散在不同社交网络的行为数据，从而准确、全面地理解用户。用户建模包括了很多方面，包括人口属性（如年龄、性别、婚姻状况和职业等）、兴趣属性（如政治、技术、音乐和运动等）、社交网络状态、流动模式、消费模式及情感倾向等。其中，人口属性记录了基本和本质的用户信息，并构成了最基础的维度来建立一般的用户模型，而兴趣属性有效直观地反映了用户喜好，二者都被广泛地用在实际信息服务中。基于跨社交媒体网络，笔者在用户人口属性建模和兴趣属性建模两方面进行了探索，下面逐一进行介绍。

## 2 跨社交媒体网络人口属性建模

近些年，很多研究根据用户的社交媒体行为对他们的人口属性进行了推断[1-6]，其中大部分研究关注提升特征性能和模型或者利用外部信息和知识。例如，Rao等人[2]利用社会语言学特征和*n*元模型，根据用户在Twitter上的行为推断他们的人口属性，包括性别、年龄和籍贯。Fang等人[4]挖掘不同人口属性之间的潜在关系，提出了一个多任务学习框架在Google+上进行关联属性推断。然而，据笔者所知，一个关键的问题被忽略且尚未解决：社交媒体行为的动态性和相对稳定的人口属性之间的矛盾。如**图1**左侧所示，用户社交媒体行为明显是随着时间变化而变化的。一方面，上述的人口属性研究通常将用户不同时间的动态行为看作一个整体，这导致了用户建模中的信息丢失，从而不能获得动态行为和稳定人口属性之间的潜在关系。另一方面，用户兴趣建模的研究已经通过将用户行为分成不同时间段来估计随着时间变化的兴趣以解决动态问题[7]。在人口属性推断的背景下，考虑到人口属性（如性别、年龄、婚姻状况和职业等）是静态的或者在很长一段时间里是不变的，动态兴趣建模的方法不能直接进行应用。

笔者通过寻找用户在不同场景下的共享模式进行推断来解决这个问题。现今每个用户都同时使用多个社交媒体网络，这为笔者提供了天然的测试网络探索用户共享行为模式来进行人口属性推断。如**图1**所示，假设存在唯一的稳定人口属性解释和导致了在各种社交媒体网络中不同的动态社交媒体行为，笔者提出了一个跨社交媒体网络的人口属性推断方法来实现上述假设。具体地，笔者考虑将Google+和Twitter作为本探索中的测试社交媒体网络。用户在每个网络中都可以发布文本、图片和视频信息。具体来说，训练模型时包含两个步骤。首先针对每个社交媒体网络，根据用户社会多媒体行为建立每个用户的特征表示。接着，将已知人口属性作为监督，利用对偶投影矩阵方法挖掘同一用户不同社交媒体网络之间的共享模式，获得人口属性空间和行为特征空间之间的关系。测试时，给定在不同社交媒体网络上

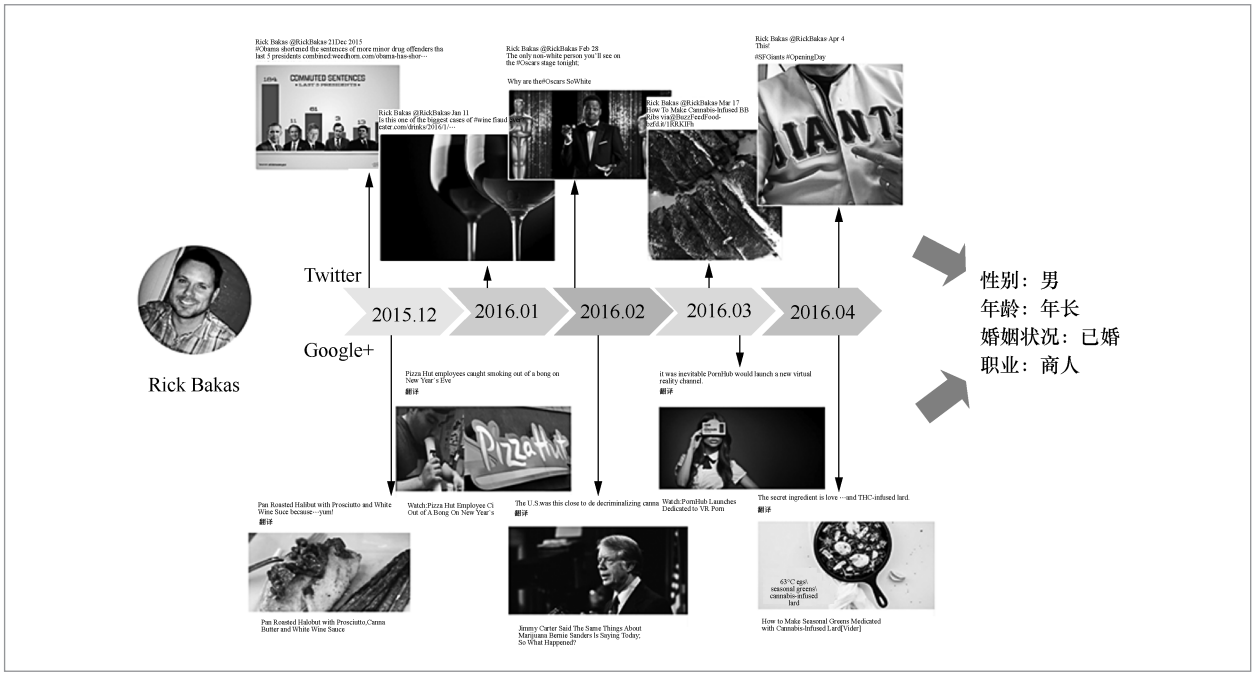



图1 对用户 Rick Bakas 根据其社会媒体行为进行人口属性推断示意

的用户行为，先提取用户特征，然后通过得到的对偶投影矩阵进行投影，最终得到用户的人口属性。

## 2.1 用户多媒体行为的特征提取

用户行为中包含大量的文本内容，可以反映用户信息。这里笔者进行了词干提取和去除停用词，并去掉了在整个文本中出现词频小于15次的词。为了减少特征表示的维数，进一步采用了基于熵的方法对每种属性选择最具有识别力的词。其基本的思路是计算每个词的互信息熵，并选取熵值最高的10 000个词。最后采用TF-IDF方法对特征进行重新加权，得到了用户的文本特征表示。

用户的多媒体行为除了包含文本内容外，还有很多的图片内容。同时考虑图片特征，可以更进一步地表示用户行为。笔者采用了广泛使用的在ImageNet（图像识别最大数据库）上训练的VGG16模型，对每一张图片根据全连接层提取了1 000维视觉特征。由于用户通常发表超过一张的图片，所以对图片表示采用最大池的方法，得到了每个用户的1 000维聚合特征向量。

最后，笔者连结文本和图片特征，获得了每个用户的社交行为特征表示。同时，分别对每个社交媒体网络都采取同样的特征提取，最终获得每个网络的每个用户的社交行为特征表示。

## 2.2 对偶投影矩阵提取

对于每个社交媒体网络，假设用户行为特征空间和人口属性空间的关联可以用投影矩阵$W$表示。因此，用户的人口属性表示$s_u$可以通过其社交行为特征$f_u$直接投影进行推断。这个假设可以用计算式表示为：$f_u=Ws_u$。笔者的工作是通过观察训练集用户的社交行为特征和他们相应的人口属性集学习投影矩阵$W$，可以通过解下面的优

化问题来实现：

$$\min_{W,S}\|F-WS\|_F^2+\lambda_1\|S-A\|_F^2+\lambda_2\|W\|_F^2 \quad (1)$$

其中，$F=[f_1,f_2,\cdots,f_N]$、$S=[s_1,s_2,\cdots,s_N]$分别是训练集中所有$N$个用户的社交行为特征和人口属性表示。$A=[a_1,a_2,\cdots,a_N]$是用户属性的离散表示，通过直接扩展用户标记属性作为一个串联的二值向量。在这里，将离散属性表示$A$修正为连续形式$S$，可以更好地反映用户不同属性值的相对强弱。

然而，在这个模型中，观察到的动态社交行为和相对稳定的人口属性之间的矛盾并没有被考虑。为了解决这个问题，笔者的方法最基本的前提是寻找不同社交网络上的大量用户行为的共享模式。因此，进一步修正式（1）中连续属性表示$S$为两个社交网络的一个共享因子，得到了下面的目标函数：

$$\min_{W^p,W^q,S}\|F^p-W^pS\|_F^2+\|F^q-W^qS\|_F^2+\lambda_1\|S-A\|_F^2+\lambda_2\|W^p\|_F^2+\lambda_3\|W^q\|_F^2 \quad (2)$$

其中，$F^p$、$F^q$分别是两个社交网络上所有$N$个用户的社交行为特征，$W^p$、$W^q$分别是两个社交网络的对偶投影矩阵，$\lambda_1$、$\lambda_2$和$\lambda_3$是3个正则化参数。这样求出的属性表示$S$利用了不同的社交网络，可以反映一些稳定行为模式。

考虑到目标函数中有很多变量，笔者采用一个等价的算法来寻找$W^p$、$W^q$和$S$的最优解，主要思想是固定其他变量最小化目标函数求一个变量，不断迭代更新直到收敛或最大迭代次数，最终获得笔者所要求解的对偶投影矩阵。

### 2.3 用户人口属性推断

根据求出的对偶投影矩阵$W^p$和$W^q$，给定一个新用户，已知他的社交行为特征$f^p$和$f^q$，可以估计他唯一的人口属性，过程如下：

$$s^*=\min_s\|f^p-W^ps\|_F^2+\|f^q-W^qs\|_F^2 \quad (3)$$

除此之外，当得到了投影矩阵，同样可以通过一些用户在单一社交网络中的社交行为数据简单地进行人口属性的粗略推断。该过程的优化函数如下：

$$s^*=\min_s\|f-Ws\|_F^2 \quad (4)$$

其中，$f$和$W$分别是某个社交网络的用户社交行为特征和投影矩阵。

已知每个用户推断得到的用户属性表示$s$，它的每一项对应了某种属性的某一属性值对应的得分，将每种属性的属性值对应的得分进行排序，选择得分最高的属性值作为该种属性的最终推断结果。

将两个流行的社交网络Google+和Twitter作为测试网络。通过Google+上用户分享的其他网络账号，笔者建立了包含1 478个共同用户的集合，并下载了他们最近发表的2 000条帖子（包括文本和图片）和用户的资料。研究其性别、年龄、婚姻状况和职业4个人口属性，以准确率为评价指标，比较了对偶投影矩阵提取（CPME）方法和投影矩阵（PME）方法、支持向量机（SVM）方法分别在两个社交媒体网络中的属性推断结果。**表1**为笔者所提方法在4个人口属性推断中与其他技术的对比结果，从**表1**可知，尽管在单网络下的投影矩阵方法的推断准确率不如支持向量机方法的推断准确率，但是对偶投影矩阵提取方法利用了丰富的跨社交网络用户数据，有效地提高了用户人口属性推断准确率。这同时说明了对偶投影矩阵提取方法可以有效地解决用户动态行为数据和相对稳定的人口属性之间的矛盾。

**表2**为对偶投影矩阵提取方法在4个人口属性推断过程中利用不同用户数据的设置的推断准确率。分别给定用户单独在Google+上的数据、单独在Twitter上的数据和两个网络上所有的数据。即使只有一

表1 不同方法的人口属性推断准确率

| | 性别 | 年龄 | 婚姻状况 | 职业 |
|---|---|---|---|---|
| SVM_Google+ | 0.747 37 | 0.693 33 | 0.583 01 | 0.391 04 |
| SVM_Twitter | 0.760 03 | 0.680 00 | 0.584 92 | 0.372 58 |
| PME_Google+ | 0.742 45 | 0.654 04 | 0.580 92 | 0.381 61 |
| PME_Twitter | 0.755 21 | 0.640 84 | 0.582 36 | 0.366 81 |
| CPME | 0.779 90 | 0.716 04 | 0.600 84 | 0.420 93 |

表2 本方法不同设置下的属性推断准确率

| | 性别 | 年龄 | 婚姻状况 | 职业 |
|---|---|---|---|---|
| Google+ | 0.751 11 | 0.706 36 | 0.592 82 | 0.393 50 |
| Twitter | 0.771 32 | 0.708 94 | 0.590 27 | 0.389 14 |
| 跨网络 | 0.779 90 | 0.716 04 | 0.600 84 | 0.420 93 |

个网络的数据，对偶投影矩阵提取方法的推断准确率仍要高于其他方法的推断准确率，因为在对偶投影矩阵提取的过程中已经得到了潜在的不同平台之间的稳定关联。同时，给定更多的用户数据可以得到更高的用户属性推断准确率。因此，基于跨社交媒体网络的用户人口属性建模可以解决动态的用户行为和相对稳定的人口属性之间的矛盾，有效地提高人口属性推断的准确率。

## 3 跨社交媒体网络兴趣属性建模

特定的社交媒体服务一般都是在单一社交媒体网络进行。例如，YouTube上的视频推荐服务已经成为引导用户从大量的视频中找到自己感兴趣的视频的一种重要方式[8]。但基于单一网络的解决方法存在着一定的局限性：单个网络上可以利用的用户数据往往不足以全面地理解用户兴趣和有效地捕捉不断改变的用户喜好。因此，笔者利用了用户分散在多个不同社交网络的数据，帮助预测用户在YouTube上的兴趣画像和视频偏好，并设计了一种统一的视频推荐解决方案，提升个性化推荐服务的效果。

统一的视频推荐方案致力于解决以下3个问题。

（1）新用户问题

当一个新用户注册到YouTube网站并且刚开始使用相应推荐服务时，系统没有关于该用户对视频兴趣的任何了解。一般而言，对新用户是利用有限的注册信息进行用户建模[9]，或是直接作为平均用户对待，并对其推荐最热门的相关物品[10]。

（2）冷启动问题

冷启动问题是指由于缺乏足够的用户初始数据，推荐系统无法提供精准推荐的相关情形。笔者用轻量用户表示那些仅有很少历史行为记录的用户。目前对轻量用户进行个性化推荐的方法包括利用用户的内容信息（如用户简介资料以及标注信息）进行基于内容的推荐[11,12]以及利用已知的社交关系数据预测用户偏好并启动推荐系统[13]等。

（3）数据稀疏性问题

在典型的推荐系统中，绝大多数用户没有机会浏览或评价大部分物品，因此用户—物品交互矩阵往往非常稀疏。这在

具有较高物品—用户比例的系统中尤为严重，如拥有超过20亿视频的YouTube网站。目前已经有不少工作专门针对减轻数据稀疏性问题，如用默认值填充缺失的用户—物品记录[13]，利用潜在因子模型将用户和物品投影到公共的低维子空间，可以捕捉用户—物品交互行为背后的潜在结构[14]以及通过传播或者迭代模型发现用户间的高阶关联[15]等。

这3个问题一直是推荐系统领域最经典的问题，受到了广泛的关注，但同时处理上述3种问题的统一解决框架还没被研究过。不同于大多数工作致力于更好地利用目标网络中的数据，笔者考虑利用其他辅助网络上的丰富的用户数据。笔者利用来自Twitter辅助网络的更多用户数据，介绍一种简单的解决框架同时处理所有上述提到的问题，有效地帮助3种典型的YouTube用户。具体地说，对于新用户，通过分析用户在Twitter网络上的推文活动，估算他们在YouTube上的兴趣画像，基于此给出一个初始化的视频推荐列表；对轻量用户，通过整合来自Twitter辅助网络的信息和用户在YouTube网络已有的部分信息来装载推荐引擎；对重度用户，通过进一步降低数据稀疏性，可以提供给他们更有效的推荐。辅助信息有助于计算用户间的关联性。

笔者的整体解决框架如**图2**所示，该框架由两个阶段构成，即辅助社交网络数据迁移与跨社交网络用户行为整合。在第一阶段，辅助网络和目标网络用户行为间的关联被嵌入一个转移矩阵中，通过该转移矩阵，用户Twitter上的推文活动可以被映射到YouTube的一个潜在用户空间。利用学习得到的转移矩阵，可以通过转移用户的推文历史行为大致估算用户的兴趣画像，并得到该用户在YouTube上的视频偏好。对于新用户，推荐系统已经可以直接利用转移得到的兴趣画像生成推荐结果。在第二阶段，以转移得到的兴趣画像为先验，进一步介绍一种基于正则约束的方法来整合两种不同的用户数据源。此外，加入一个权重矩阵，根据用户可得到的YouTube行为数据自适应地调节不同源的整合权重。以这种方式，得到的轻量用户和重度用户的兴趣画像同时考虑了Twitter上的用户推文活动以及历史的YouTube视频行为交互情况。然后得到的用户模型可以直接用来生成相应的推荐结果。

### 3.1 辅助社交网络行为迁移

笔者通过测试共同用户在Twitter上的辅助数据如何被转移到其对YouTube上的兴趣画像来发现相应的关联模式。笔者提出的解决方案是基于有约束的矩阵分解方法的。在推荐系统中，矩阵分解模型将用户和物品投影到一个潜在的因子空间，其中用户—视频交互被模拟为二者的内积。因此，用户在YouTube上的兴趣画像就是用户的潜在因子表示$u$，已经嵌入相应的

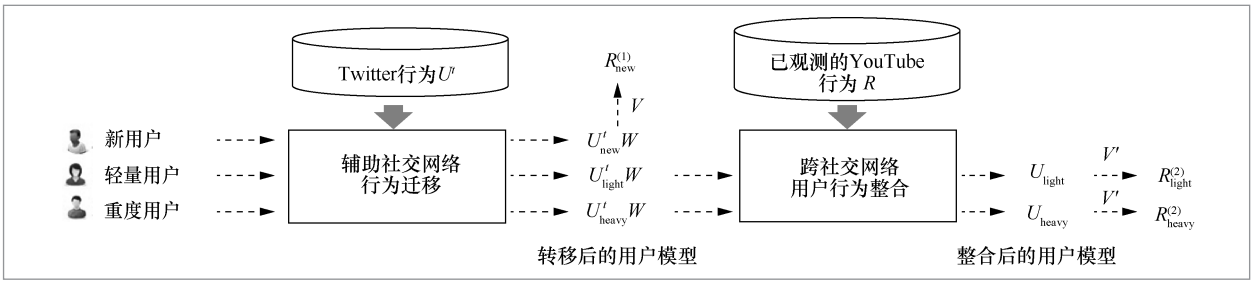



图 2 基于跨网络用户行为的统一视频推荐整体解决框架示意

用户—视频交互矩阵$R$中。首先，为了表示用户在Twitter上的推文活动，视每个用户的推文历史为一个文档，将标准的LDA主题模型应用到所有Twitter用户构成的文档集上。结果，每个用户可以被表示为一个主题分布$u^t$。对每个跨网络共同用户，假设存在一个转移矩阵$W$蕴含着从该用户的Twitter主题分布$u^t$向其由交互矩阵中提取的YouTube用户兴趣画像$u$进行映射的映射关系。因此，转移辅助数据的任务变为利用跨网络共同用户在Twitter和YouTube上的以观测行为学习相应的转移矩阵$W$。笔者利用用户Twitter主题分布$u^t$和转移矩阵$W$替代用户兴趣画像$u$，构建有约束的矩阵分解模型。其中，为了防止过拟合，仅基于已观测到的交互数据去发现潜在数据结构[14,16]。同时，笔者融合内容信息作为正则约束[11,17]，使得具有相似内容信息的视频在得到的潜在因子空间具有相似的表示，可以有效地减少数据稀疏性的影响。最终，通过求解该模型，可以得到相应的转移矩阵$W$和视频的潜在因子表示$V$，捕捉到用户在辅助网络和目标网络行为的关联性。那么，给定任意测试用户及其Twitter主题分布$u^t$，就可以估算该用户在YouTube上的兴趣画像及视频的偏好。

在**图3**中，笔者模拟了3类YouTube用户，给出了相应的简单示例。通过$W$转移，预测用户在包含科技、游戏和体育等潜在视频主题上的用户兴趣（×、√、√√分别表示“不喜欢”“喜欢”和“非常喜欢”）。进一步与$V$相乘，可以发现用户在特定视频上的相关偏好。因此，即使在目标网络没有任何可利用的行为记录的情况下，仍然可以通过转移辅助网络的数据来建立一个初始化的用户兴趣画像。这实际上解决了统一视频推荐问题下的第一种情形：新用户。

### 3.2 跨社交网络用户行为整合

对轻量用户和重度用户，他们在目标网络已经有部分观测到的行为。直接将估算得到的视频偏好与目标网络已观测到的行为进行聚合往往不太现实，因为这两个网络上的行为可能存在不一致性甚至相互矛盾。因此，同时考虑已观测到的YouTube用户—视频矩阵$R$和转移得到的用户兴趣画像$U^tW$来更新轻量用户和重度用户的兴趣画像，作为他们的潜在用户表示$U$。

笔者视Twitter转移得到的用户兴趣画像$U^tW$为整合后的用户兴趣画像$U$的先验。这可以从两方面进行解释：对在目标网络具有稀疏的观测行为的用户，Twitter转移

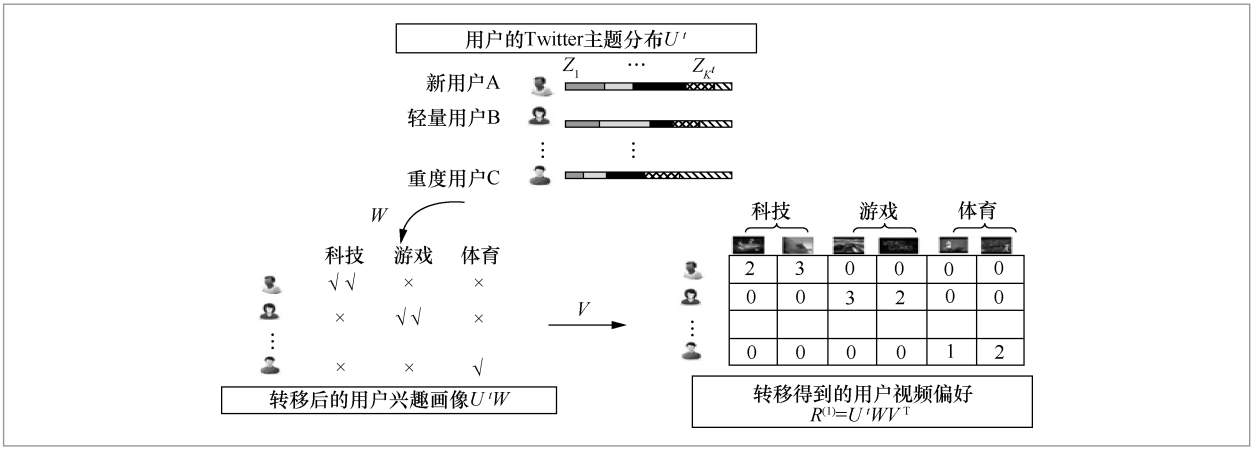



图 3 辅助社交网络行为迁移的示例

得到的用户兴趣画像$U^tW$可以作为整合用户兴趣画像的一个很好的表示，即$U$应该与$U^tW$类似，这实际上对应着冷启动问题。求得的转移矩阵$W$定义了用户所在的一个潜在空间，这有助于计算用户间的关联性，并能被利用以减轻目标用户—视频矩阵$R$中的稀疏性问题。因此，需要利用辅助社交网络行为迁移后得到的视频潜在表示$V$和用户兴趣画像$U^tW$，正则化约束更新学到的视频潜在因子$V'$和用户兴趣画像$U$可以更好地拟合观测到的用户—视频矩阵$R$，并且更新的视频潜在因子$V'$尽可能接近已求得的视频潜在因子$V$，而Twitter转移得到的用户兴趣画像$U^tW$和更新后的用户兴趣画像$U$之前的差也按照一定的权重进行约束，最终得到最优的更新后的用户兴趣画像$U$和视频潜在表示$V'$。对每个测试的轻量用户或者重度用户，都可以计算其在YouTube视频上的偏好。

在**图4**中，接着**图3**的示例，进一步展示了轻量用户和重度用户的用户兴趣画像在整个过程中如何更新，可以看出根据YouTube上观测到的视频行为，转移的用户兴趣画像和视频表示被进一步改进（修改的矩阵元素加粗突出）。最终得到的用户视频偏好既考虑了Twitter网络的辅助信息，又很好地拟合了目标网络已观测到的用户行为。因此，笔者解决了统一视频推荐问题中的剩下的两种情形：轻量用户和重度用户。

## 4 结束语

社交媒体中的异构和多源问题是深度利用社交媒体大数据的关键。随着人们对不同社交媒体服务的深入和跨网络共同用户发现技术的成熟，大量可获得的共同用户对应关系可以作为跨社交媒体网络分析的桥梁，有效地连接各社交媒体网络中的独立异构数据，并予以综合利用。同时，用户作为社交媒体服务的中心，为了更好地进行个性化信息服务，迫切地需要进行用户建模，达到全面准确地理解用户的目的。

本文从跨社交媒体网络用户人口属性建模和兴趣属性建模两方面讨论了对跨社交媒体网络中多源异构数据的综合利用。因为用户在不同社交媒体网络中的行为都在一定程度上反映了其属性，所以可以通

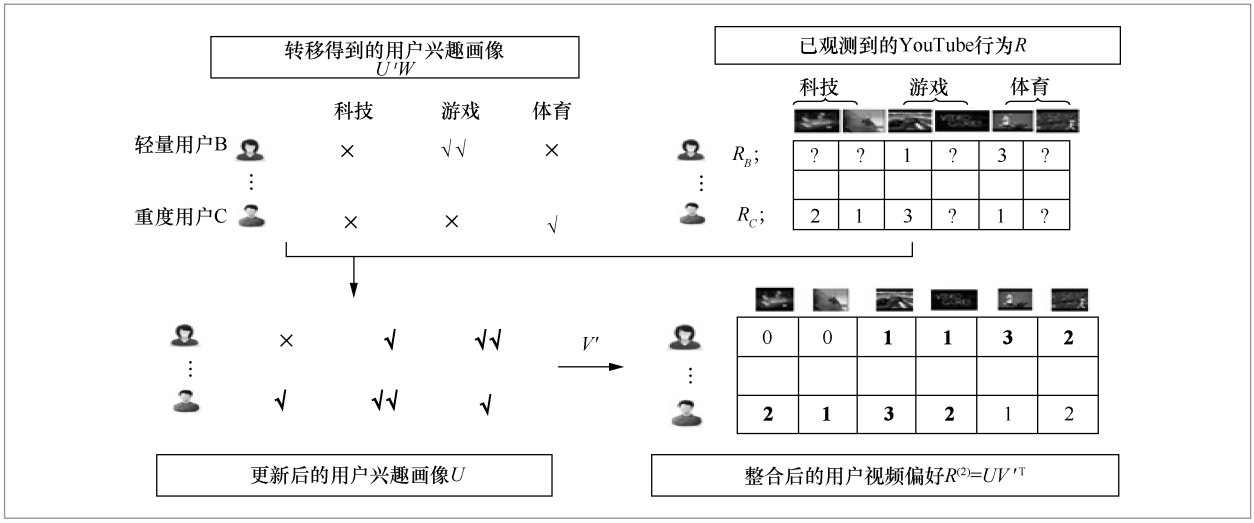



图 4 跨社交网络用户行为整合的示例

过整合用户在不同社交媒体网络上的行为信息进行协同分析，有效地解决单网络的行为稀疏性和内容单一性等问题。未来，笔者将从以下几个方面对跨社交媒体网络工作展开进一步地研究：更进一步分析不同网络之间的关联和区别，充分挖掘数据所蕴含的信息；在用户人口属性建模和兴趣属性建模基础上，着眼于更多应用，更好地利用大数据来服务于用户。

## 参考文献:


[1] ZHELEVA E, GETOOR L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles[C]//The 18th International Conference on World Wide Web, April 20-24, 2009, Madrid, Spain. New York: ACM Press, 2009: 531-540.

[2] RAO D, YAROWSKY D, SHREEVATS A, et al. Classifying latent user attributes in twitter[C]//The 2nd International Workshop on Search and Mining User-Generated Contents, October 30, 2010, Toronto, Canada. New York: ACM Press, 2010: 37-44.

[3] PENNACCHIOTTI M, POPESCU A M. Democrats, republicans and starbucks afficionados: user classification in twitter[C]//The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 21-24, 2011, San Diego, CA, USA. New York: ACM Press, 2011: 430-438.

[4] FANG Q, SANG J, XU C, et al. Relational user attribute inference in social media[J]. IEEE Transactions on Multimedia, 2015, 17(7): 1031-1044.

[5] CHEN X, WANG Y, AGICHTEIN E, et al. A comparative study of demographic attribute inference in twitter[C]//The 9th International AAAI Conference on Web and Social Media(ICWSM), May 26-29, 2015, Oxford, UK. [S.l.:s.n.], 2015: 590-593.

[6] HUANG Y, YU L, WANG X, et al. A multi-source integration framework for user occupation inference in social media systems[J]. World Wide Web, 2015, 18(5): 1247-1267.

[7] SANG J, LU D, XU C. A probabilistic framework for temporal user modeling on Microblogs[C]//The 24th ACM International on Conference on Information and Knowledge Management, October 19-23, 2015, Melbourne, Australia. New York: ACM Press, 2015: 961-970.

[8] DAVIDSON J, LIEBALD B, LIU J, et al. The YouTube video recommendation system[C]//The Fourth ACM Conference on Recommender Systems, September 26-30, 2010, Barcelona, Spain. New York: ACM Press, 2010: 293-296.

[9] DEGEMMIS M, LOPS P, SEMERARO G. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation[J]. User Modeling and User-Adapted Interaction, 2007, 17(3): 217-255.

[10] JANNACH D, ZANKER M, FELFERNIG A, et al. Recommender systems: an introduction[M]. [S.l.]:Cambridge University Press, 2010.

[11] PAZZANI M J, BILLSUS D. Content-based recommendation systems[M]//The Adaptive Web. Berlin: Springer Berlin Heidelberg, 2007: 325-341.

[12] ZHANG Z K, LIU C, ZHANG Y C, et al. Solving the cold-start problem in recommender systems with social tags[J]. Europhysics Letters, 2010, 92(2): 28002-28007.

[13] DESHPANDE M, KARYPIS G. Item-based top-n recommendation algorithms[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 143-177.

[14] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//The 14th

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, Las Vegas, Nevada, USA. New York: ACM Press, 2008: 426-434.

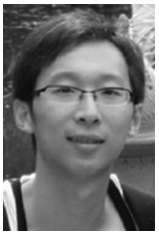
[15] HUANG Z, CHEN H, ZENG D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 116-142.

[16] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization[C]// The 25th Annual Conference on Neural Information Processing Systems(NIPS), December 12-17, 2011, Granada, Spain. [S.l.:s.n.], 2011: 880-887.

[17] BALABANOVIĆ M, SHOHAM Y. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66-72.

作者简介

**项连城**(1992-),女,中国科学院自动化研究所硕士生,主要研究方向为社交多媒体分析与挖掘。



**桑基韬**(1985-),男,博士,中国科学院自动化研究所副研究员,主要研究方向为社会媒体分析、多媒体检索、数据挖掘。

**徐常胜**(1969-),男,博士,中国科学院自动化研究所研究员,中国科学院大学博士生导师,主要研究方向为多媒体分析/索引/检索、模式识别、计算机视觉。

**收稿日期:** 2016-08-12

**基金项目:** 国家自然科学基金资助项目(No.61432019, No.61225009, No.61303176)

**Foundation Items:** The National Natural Science Foundation of China(No.61432019, No.61225009, No.61303176)