DOI: 10.11992/tis.201706036
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1250.016.html

基于文本扩展模型的网络视频聚类方法

刘璐[1,2], 贾彩燕[1,2]
（1. 北京交通大学 交通数据分析与挖掘北京市重点实验室, 北京 100044; 2. 北京交通大学 计算机与信息技术学院, 北京 100044）

摘　要: 随着视频分享网站的兴起和快速发展, 互联网上的视频数量呈爆炸式增长, 对视频的组织及分类成为视频有效使用的基础。视频聚类技术由于只需要考虑视频数据内在的簇结构、不需要人工干预, 越来越受到人们的青睐。现有的视频聚类方法有基于视频关键帧视觉相似性的方法、基于视频标题文本聚类的方法、文本和视觉多模态融合的方法。基于视频标题文本聚类的视频聚类方法由于其简便性与高效性而被企业界广泛使用, 但视频标题由于其短文本的语义稀疏特性, 聚类效果欠佳。为此, 本文面向社会媒体视频, 提出了一种社会媒体平台上视频相关多源文本融合的视频聚类方法, 以克服由于视频标题的短文本带来的语义稀疏问题。不同文本聚类算法上的实验结果证明了多源文本数据融合方法的有效性。

关键词: 网络视频聚类; 共点击视频; 相关查询词; 文本聚类
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2017)06−0799−07

中文引用格式: 刘璐, 贾彩燕. 基于文本扩展模型的网络视频聚类方法[J]. 智能系统学报, 2017, 12(6): 799–805.
英文引用格式: LIU Lu, JIA Caiyan. Web video clustering method based on an extended text model[J]. CAAI transactions on intelligent systems, 2017, 12(6): 799–805.

Web video clustering method based on an extended text model

LIU Lu[1,2], JIA Caiyan[1,2]
(1. Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China; 2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With the rapid rise and development of video sharing websites, there has been an explosive increase in web videos on the Internet. Effective organization and classification are necessary for the valid use of such videos. Video clustering technology has gained increasing popularity because it considers the internal cluster structure of video data, and no manual intervention is necessary. There are many video clustering algorithms in existence, such as those based on the visual similarity of key frames, text clustering of video titles, and multi-model fusion by integrating text and visual features. The video clustering method based on the text clustering of titles has become a widely used method in business because of its simplicity and efficiency. However, it performs poorly due to the semantic sparsity of short titles. Therefore, this paper proposes a video clustering method with related text fusion from multiple sources on social media platforms to overcome the semantic sparsity of short text. The experimental results on different text clustering algorithms demonstrate the effectiveness of this method.

Keywords: web video clustering; co-click videos; relevant inquiry word; text clustering

伴随着网络多媒体技术不断的应用和发展, 网络视频作为一种重要的传播媒介, 凭借其丰富多彩的内容和便捷的传播形式深受广大网络用户的喜爱。网络视频分享应用及网站在其不断发展中逐渐形成了独特的特点, 在这些网站上用户可以申请上传视频, 也可以从海量视频中选择观看自己感兴趣的视频。在社会媒体平台上广泛存在的视频具有如下特点:

收稿日期: 2017−06−09.　**网络出版日期:** 2017−11−09.
基金项目: 国家自然科学基金项目 (61473030).
通信作者: 贾彩燕. E-mail: cyjia@bjtu.edu.cn.

1) 网络视频规模大, 上传更新速度快;

2) 通常以用户制作内容和专业影视内容并存为主要特色, 内容涉及电影、音乐、科技、汽车、亲子、时尚、体育、财经、资讯等多个主题;

3) 不同于传统图像或文本, 网络视频蕴含丰富的多模态信息: 视频的视觉图像、标题、描述、标签等文本信息, 观看用户的评论, 上传用户信息等;

4) 网络视频时长不等, 剪辑水平不等。

如此规模宏大、类别众多的网络视频在丰富互联网信息, 满足用户需求的同时, 也给视频数据的组织和管理带来了严峻的挑战。传统的解决办法是视频上传者或者网站管理员人工标注视频的类别, 然而这一方法耗时耗力, 并且容易受到人为不稳定的主观影响。有学者提出利用基于监督或半监督的视频自动分类技术[1-2], 但是大部分分类方法仍旧需要充足的高质量训练样本。因此, 研究人员尝试利用无监督的机器学习模型——聚类来进一步组织网络视频数据, 提出了多种基于文本、视觉及多模融合的视频聚类方法[3-8]。另外, 现有的视频搜索引擎的搜索结果通常按与搜索词的相关度进行排列, 其中可能包含重复或相似视频, 为了给用户提供更可观、更易理解的搜索体验, 因而提出了基于文本特征、视觉特征或多特征融合的视频搜索结果聚类方法[3-4], 便于对视频按主题进行推荐。

单纯基于视频文本 (如标题) 的聚类可以利用已有文本聚类的成熟技术, 考虑视频的语义特征[4], 具有简单、易用等特点。但视频标题作为短文本, 存在文本特征高维语义极度稀疏等问题。单纯使用视觉相似性的视频聚类方法[5]由于图像语义理解的复杂性, 存在维度高、复杂度高、聚类效果欠佳等问题。网络视频作为一种社会化媒体数据, 形式丰富、模态多样, 包含视频视觉内容、描述单一主题的标题文本、播单信息[6]、上传及观看用户等。利用文本和视觉等多模态融合的视频聚类方法[3, 7], 可以增强单一方法视频聚类的有效性, 但存在时间复杂度高、最优加权方案不确定等问题。鉴于文本挖掘技术的成熟性和易用性, 本文着眼于利用社会媒体上丰富的文本信息, 以改善现有短文本方法的高维、语义稀疏问题, 实现高效的视频主题聚类, 提出了以视频标题、相关查询词、共点击视频标题等多类短文本信息融合的视频主题聚类方法, 并以优酷视频网站 (http://www.youku.com) 真实数据为例, 验证了本文方法的有效性。

1　相关工作

1.1　网络视频聚类

目前, 网络视频聚类已经取得一些研究成果, 主要有基于视觉图像的方法、基于文本特征的方法、基于视频其他信息 (如播放列表) 的方法以及多模融合的方法。

Liu 等[5]使用从视频帧序列提取的全局视频签名特征 (ViSig) 来表征视频, 提出了基于视觉图像相似度的视频搜索结果聚类算法。Nguyen 等[4]从网络视频的文本元数据 (视频标题、标签和描述) 角度出发, 提出了基于 WordNet 知识库的文本语义相似度计算方法, 可以有效提高视频搜索结果的聚类效果。文献[3, 7]是基于视频的多模态特征进行主题聚类。Hindle等[3]集成网络视频的视觉特征和文本特征, 利用有界坐标系统 (BCS) 模型将视频的视觉特征用紧凑签名来表示, 然而在计算文本相似度时, 该方法将文本近似视为“词袋”, 直接统计两个文本共有词的个数, 忽略了文本信息的高级语义知识。Huang 等[7]针对网络视频包含的丰富信息, 分别计算低层视觉图像特征、高层语义特征和文本特征的相似度, 然后将各个模态融合计算网络视频的实际相似度, 并引入近邻传播的聚类算法进行网络视频聚类分析。Zhang 等[8]从视频语音转录文本和视觉概念识别两方面提取视频特征, 提出二部图聚类算法, 表明在多源数据关联挖掘方面效果要优于常规谱聚类。研究学者还利用网络视频丰富的社会媒体信息展开研究。Kamie 等[6]提出 PVClustring 方法, 构建视频–播单关联矩阵, 采用重复二分的 kmeans 方法进行视频聚类; Zhang 等[9]利用 YouTube 的共观看视频进一步改进网络视频分类系统的性能。

由此可见, 研究人员尝试利用不同的理论方法来解决视频聚类问题。但是当前的研究中, 使用低层视觉特征仅在识别雷同视频时有较好的效果, 同时由于视频图像语义理解的复杂性、以及海量、高维、语义不清晰等特点, 现有的利用视频信息的多模态融合方法对视频聚类的效果还不能满足实际需求, 缺乏有效的主题聚类方法。

1.2　短文本聚类

伴随着 Web2.0 时代, 短文本数据在互联网上应用日益增多, 短文本聚类的相关工作也取得了很大的进展, 研究者们尝试利用很多方法来改进短文本语义分析与处理, 大体上分为两类。一类是挖掘短文本自身内容构建特征空间, Yin 等[10]提出了应用吉布斯抽样的狄利克雷混合模型算法 (GSDMM), 在聚类过程中可以自动推断出类别数量并快速收敛, 能很好地适应短文本高维稀疏的状况; Yan 等[11]结合上下文语义相关性来建立词项关联矩阵, 避免使用了短文本中高维稀疏的词文档矩阵, 然后应用

对称非负矩阵分解算法获取词项–主题矩阵进而推断每个文档的主题。

另一类是利用外部知识库来扩展短文本表示, Sahami 等[12]通过利用网络搜索结果扩展短文本内容, 在扩展的基础上计算文本间相似度; Yih[13]在 Sahami 的基础上通过计算词出现的加权内积而不是 TFIDF, 并引入了一个学习过程来提高相似性度量的准确性; Banerjee 等[14]利用从短文本中提取出的字符串检索维基百科中最相关的前 10 个文档, 并用这些文档的标题扩充每个短文本文档的表示, 再对短文本进行聚类; Gabrilovich 等[15]提出了一种显示语义分析, 将每个短文本映射到最相关的维基百科和 ODP(开放目录项目) 的本体概念, 用概念向量扩充传统的词袋模型表示; Hu 等[16]同时采用内部特征和外部特征 (维基百科和 WordNet)来对短文本文档进行扩充, 提出了一个分层的三级结构来解决原始短文本的数据稀疏问题; Hotho 等[17]将 WordNet 集成到文本聚类的过程, 在 Reuters语料库的实验结果显示了它的有效性; Song 等[18]利用开放的网页构建了一个概率化知识库, 进而来推断短文本文档中的概念表示, 然后再进行聚类。这些方法已经被证明能有效地提高短文本聚类, 然而利用搜索引擎的短文本扩展方法时间复杂度高, 在利用外部知识库扩展的方法中, 寻找合适的外部源也十分重要, 但是由于互联网的自由开放性, 网络视频的标题文本一般由用户上传视频时自己填写, 容易出现新兴词汇和网络用语, 语言表达方式和其他长文本文档有着较大的差异, 盲目地扩充可能会影响原短文本的语义。

2　网络视频表示

随着视频分享网站的不断应用, 网络视频不再仅仅是单一的视频结构, 而是作为一种丰富的多媒体信息包含了多源数据。网络视频的播放页面, 不仅包含具体的视频内容, 还包含标题、描述、标签等用户提供的文本信息, 以及用户之间评论、点赞、收藏等社交互动行为。在已有的研究工作中, 文献[7]的实验表明利用标题等文本特征在视频聚类上有较好的效果。同时, 在实际工业应用中, 利用视频的图像特征进行视频表征时, 存在图像存储占空间、时间复杂度高, 只适用于短视频或视频画面内容较集中的视频等多种问题。本文研究使用视频的多源文本信息来更准确地表示视频, 包括视频标题、视频相关查询词、共点击视频标题, 利用这些信息进行聚类能够从语义层次上有效识别视频聚簇。仅使用标题短文本进行特征表示时由于字数较少, 不同文本间缺乏足够的词共现信息, 因而存在高维稀疏、特征模糊、语义不清晰问题, 而视频的相关查询词与该视频往往语义相关, 共点击视频的视频标题和该视频标题的词汇也语义相似, 利用这些信息可以进一步扩展文本内容, 丰富文本表示。因此, 本文提出一种多源数据下文本扩展模型进行网络视频表示, 为聚类研究做好准备工作, 以改善短文本高维稀疏的问题, 有效实现主题聚类。本文方法的框架如图 1 所示。

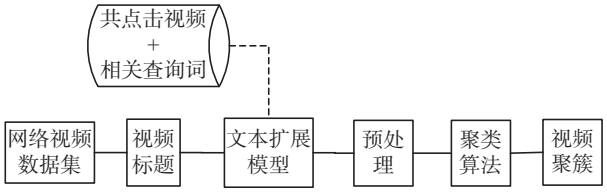


图 1　基于文本扩展模型的网络视频聚类方法框架

Fig. 1　Framework for web video clustering based on extended text model

2.1　多源数据

2.1.1　网络视频标题

视频的元信息 (meta data) 包括标题、描述和标签等文本信息, 准确的文字描述可以提供最直接有效的视频语义信息, 不同于标签具有很大的噪音、描述只在较少视频中出现, 标题作为每个视频都具备的一种短文本信息, 可以很好地概括视频的语义内容, 是描述视频的一项重要的文本特征。

2.1.2　相关查询词

每个视频会对应一系列相关的查询词, 这些查询词和该视频的标题信息通常在语义上有很大的相关性, 描述相关的视频内容。

2.1.3　共点击视频

用户在视频网站的一个会话访问过程中, 观看的视频内容通常与用户当时的兴趣密不可分。因此, 根据全网用户的点击观看行为可以将每个视频和一系列共点击视频相关联, 从一定程度来讲, 这一系列共点击视频和该视频更倾向于内容相关, 它们的标题更倾向于语义相关。这类比于文档中的词汇关系, 如果两个词语在很多文档中都频繁共同出现, 则这两个词语有很大可能是语义相关的。

2.2　文本扩展模型

针对网络视频的多源数据, 本文构建了一个多源数据下的文本扩展模型, 将短文本扩展成较长的文本以丰富语义内容, 强化词语的共现特征。针对每个网络视频, 我们不仅可以获得视频标题 (T_1), 还可以得到该视频的相关查询词 (T_2), 以及该视频的共点击视频所对应的视频标题 (T_3)。利用文本 T_2 和文本 T_3 分别去扩展原视频标题即文本 T_1, 构

成新的长文本作为该视频的文本表示。同时，利用 T_1、T_2 和 T_3 合并构成长文本进行实验对比。

在表 1 中，给出了数据集中部分视频标题及它们相应的扩展文本。例如第一个视频标题通过扩展可以补充“乒乓球”、“直拍”等词汇，丰富了原视频短文本的语义。

表 1　模型示例

Table 1　Model example

视频标题	相关查询词	共点击视频标题
王皓苦练拉球 恩师吴敬平陪练	王皓 乒乓球 比赛	第八期_直拍正手拉下旋球
徕卡LEICA T(Type 701)	徕卡t leica t camlogic相机逻辑camlogic相机逻辑 htc vive佳能80d 80d pro6	适马Sigma DP2M相机 评测
oppor7plus魅族mx5 小米note魅族	科技美学 荣耀7评测 华为荣耀7 魅蓝metal魅族	OPPOR7拆机换屏视频教程 华为P8
魅蓝note2与苹果6plus区别	pro5 马自达mx5 努比亚z9max	开箱上手体验by三宋大国论

3　实验与分析

为了定量地比较不同模型的性能，应用多种文本聚类算法进行实验，然后从准确率和标准化互信息两个方面进行分析，进而实现网络视频的聚类效果评估。

3.1　数据集

实验数据集来源于优酷视频 (http://www.youku.com)，分为两个子集 (数据集 1 和数据集 2)，均包含视频标题、共点击视频、相关查询词等数据。数据集 1 中包含亲子、汽车和科技共 3 个类别，数据集 2 中涉及广告、搞笑、电影、体育、时尚、亲子、汽车、拍客、旅游和科技共 10 个类别，每个类下的视频数量更贴近于实际网络中的分布情况，表 2 展示两个数据集的相关统计信息。

表 2　数据描述

Table 2　Description of data sets

数据集	数据集1	数据集2
视频个数	3 839	14 150
类别个数	3	10
最大类视频个数	2 622	2 869
最小类视频个数	455	186
类平均视频个数	1 280	1 415

在得到每个视频的文本表示之后，对文本的预处理过程如下：

1) 利用 jcseg 方法对文本进行切词；

2) 过滤常用停用词，以及在视频短文本中的常用非重要词汇，如“视频”、“高清”等；

3) 过滤文档频率 df 值小于 10 的词汇；

4) 过滤包含词汇数目小于 3 的文档。

3.2　评价函数

在实验数据集中，每个视频数据都已对应一个合理的类标签，因此选取了准确率 (ACC) 和标准化互信息 (NMI) 作为评估聚类算法性能的指标。

准确率是一个普遍流行的聚类质量评价指标，指正确指派类标的文档在所有文档中所占的比例，定义如下：

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta(l_{\mathrm{ti}},p_{\mathrm{map}}(l_{\mathrm{pi}}))}{n}$$

式中：$\delta(x,y)$是指克罗内克函数，如果 $x=y$，则其输出值为 1，否则为 0；l_{ti} 是文档 T_i 的真实标签；l_{pi} 是算法得出的标签；$p_{\mathrm{map}}(l_{\mathrm{pi}})$是将 l_{pi} 映射到真实对应的标签；n 是总的文档数目。显然，准确率越大，聚类划分也就越准确。

标准化互信息用来刻画一个数据集上的聚簇划分结果和此数据集真实类标的相似程度，其定义如下：

$$\mathrm{NMI}(C,C')=\frac{-2\sum_{\mathrm{i}=1}^{K}\sum_{j=1}^{K'}n_{ij}\log\dfrac{n_{ij}n}{n_i^C n_j^{C'}}}{\left(\sum_{i=1}^{K}n_i^C\log\dfrac{n_i^C}{n}\right)+\left(\sum_{j=1}^{K'}n_j^{C'}\log\dfrac{n_j^{C'}}{n}\right)}$$

式中：C 为数据集的人工标注类标，C'为由聚类算法得出的类别结构，K 为真实的聚簇数目，K'为算法得到的聚簇数目，n_{ij} 为同时在簇 C_i 和簇C'_j中数据对象的个数，n_i^C为簇 C_i 中数据对象的个数，$n_j^{C'}$为簇C'_J中数据对象的个数。通常 NMI 值越大，表明算法得到的结果越准确。

3.3　实验方法

本文对比实验了 4 种网络视频的文本表示。

1) Title：只利用网络视频的标题 (T_1) 作为该视频的文本表示。

2) Co-query enhancement：利用网络视频的标题 (T_1) 和相关查询词 (T_2) 合并作为该视频的文本表示。

3) Co-click enhancement：利用网络视频的标题 (T_1) 和共点击视频所对应的标题序列文本 (T_3) 合并作为该视频的文本表示。

4) All enhancement: 利用 T_1、T_2、T_3 合并作为该视频的文本表示。

在以上文本扩展模型的基础上, 对比实验了 6 种文本聚类算法, 前 3 种方法是面向长文本设计的聚类算法, 后 3 种是针对短文本的聚类方法:

1) latent dirichlet allocation (LDA)[19-20]: LDA 是一种文档主题生成概率模型, 本实验采用吉布斯采样学习 LDA 参数的 Java 实现, 设置参数 α 为 0.5, β 为 0.1, 迭代次数为 200 次。

2) Kmeans++[21]: 是基于相似度的聚类模型, 在 Kmeans 算法的基础上解决了需要人为确定初始聚类中心的问题, 它选取的原则是初始聚类中心之间的距离要尽可能远。

3) GNMF[22]: 图正则化非负矩阵分解模型, 在 NMF 的基础上, 构建近邻图来考虑数据样本中的几何近邻结构。

4) TNMF[11]: 是在词关联矩阵上应用非负矩阵分解的话题模型, 分为两个连续的子过程, 即话题学习和话题推断。在词关联矩阵中应用对称非负矩阵分解方法可以有效地避免传统词–文档矩阵的稀疏性。设置参数 λ 为 1。

5) GSDMM[10]: 利用吉布斯抽样算法的狄利克雷多项混合模型, 能自动推断出集群的数量, 在集群的完整性和类内同质性之间达到平衡, 收敛速度较快。设置参数 α 为 0.1, β 为 0.1, 迭代次数为 100 次。

6) biterm topic model(BTM)[23]: 是一种双词话题模型, 该模型通过直接对文档中双词 (即共现词对) 的产生进行建模来学习话题, 避免了受短文本长度过短导致的内容稀疏性影响。

对于 LDA、GSDMM、TNMF、BTM、GNMF 等话题模型, 我们将每个文本指派到其最大隶属度所在的话题聚簇以实现聚类划分。

3.4　实验结果

在上述两个数据集中分别设置聚簇数目为 3 和 10, 并根据已有类标签分别计算准确率和标准化互信息, 对于每一个方法我们重复运行 10 次计算平均值。

实验结果如表 3 ~ 6 所示, 在 LDA、Kmeans++、GNMF 方法上, 经过文本扩展之后的视频聚类效果要普遍优于只采用视频标题进行特征表征的方法, 尤其体现在 NMI 的变化上, 原因在于通过多源数据下的文本扩充, 一定程度上丰富了文本的语义信息, 增强了词共现特征, 使长文本聚类方法的性能发挥更好。在 TNMF 方法上, co-query enhancement 模型和 all enhancement 模型的效果提高最为明显, 说明扩展相关查询词相比共点击视频标题能尽可能避免加入噪音词汇, 提高实验结果。在 GSDMM 和 BTM 两种短文本聚类方法上, 本文提出的多源数据模型效果不突出, 原因在于这两种方法更适用于语义突出的较短文本, 在利用相关查询词或共点击视频标题进行扩充后带来了不可避免的噪音文本, 影响了短文本聚类效果。然而, 在这两种方法中 NMI 值有所提高, 也印证了增加共点击信息对于类大小不平衡的视频数据有较好的提升效果。

表 3　数据集 1 算法结果 (ACC)

Table 3　Result of algorithms on data set 1 (ACC)

模型	LDA	Kmeans++	GNMF	TNMF	GSDMM	BTM
Title	0.610 9 (±0.005 7)	0.582 4 (±0.000 9)	0.618 2 (±0.000 5)	0.769 0 (±0.009 9)	0.679 3 (±0.014 5)	0.790 2 (±0.012 9)
Co-query enhancement	0.706 6 (±0.010 7)	0.512 1 (±0.000 1)	0.882 0 (±0.005 6)	0.778 0 (±0.000 1)	0.729 0 (±0.009 4)	0.781 1 (±0.005 5)
Co-click enhancement	0.693 4 (±0.003 1)	0.556 9 (0)	0.897 9 (0)	0.656 3 (±0.006 8)	0.675 9 (±0.003 6)	0.666 2 (±0.002 4)
All enhancement	0.692 2 (±0.003 1)	0.507 2 (0)	0.930 5 (0)	0.781 0 (±0.000 1)	0.668 2 (±0.005 5)	0.763 1 (±0.004 4)

同时对实验结果进行横向比较, 在数据集 1 中 all enhancement 模型在 GNMF 文本聚类方法上取得了最好的效果, 在数据集 2 中从模型优化效果来看, co-click enhancement 模型在 LDA 文本聚类方法上整体效果要好, 体现出合理利用视频的相关信息可以达到最佳的视频聚类效果, 并且在本文提出的模型之上, 长文本聚类方法的效果优于短文本聚类方法。

在 10 类真实数据集中, 精确度和标准化互信息数值普遍不高, 原因主要在于如今的互联网视频中, 视频内容众多丰富, 综合类划分之下的视频可以细分为多个具体内容, 比如体育类包含篮球、足球、搏击等多项运动, 造成聚簇特征不明显, 影响了聚类效果。

表 4 数据集 1 算法结果（NMI）

Table 4 Result of algorithms on data set 1 (NMI)

模型	LDA	Kmeans++	GNMF	TNMF	GSDMM	BTM
Title	0.361 7 (±0.000 9)	0.205 8 (±0.002 8)	0.273 9 (±0.000 2)	0.417 3 (±0.009 0)	0.441 3 (±0.008 6)	0.479 9 (±0.012 4)
Co-query enhancement	0.401 0 (±0.010 3)	0.224 7 (0)	0.567 0 (±0.005 9)	0.463 7 (±0.001 0)	0.440 0 (±0.006 8)	0.467 5 (±0.012 2)
Co-click enhancement	0.467 9 (±0.001 6)	0.277 8 (0)	0.608 2 (0)	0.380 9 (±0.008 0)	0.472 4 (±0.001 9)	0.466 8 (±0.001 7)
All enhancement	0.420 3 (±0.005 7)	0.242 2 (0)	0.674 5 (0)	0.503 8 (±0.000 5)	0.402 5 (±0.002 3)	0.464 6 (±0.004 3)

表 5 数据集 2 算法结果（ACC）

Table 5 Result of algorithms on data set 2 (ACC)

模型	LDA	Kmeans++	GNMF	TNMF	GSDMM	BTM
Title	0.389 5 (±0.000 9)	0.297 9 (±0.000 4)	0.305 3 (±0.000 1)	0.346 0 (±0.000 9)	0.492 5 (±0.001 2)	0.512 7 (±0.000 5)
Co-query enhancement	0.438 3 (±0.001 1)	0.330 1 (±0.000 7)	0.429 1 (±0.000 2)	0.366 2 (±0.002 4)	0.420 4 (±0.002 2)	0.478 2 (±0.000 2)
Co-click enhancement	0.464 6 (±0.000 8)	0.375 8 (±0.000 7)	0.417 2 (±0.000 6)	0.328 5 (±0.001 4)	0.461 4 (±0.001 9)	0.465 4 (±0.000 1)
All enhancement	0.451 7 (±0.000 6)	0.358 2 (±0.000 4)	0.438 9 (±0.000 8)	0.373 9 (±0.001 2)	0.418 2 (±0.001 5)	0.459 3 (±0.000 5)

表 6 数据集 2 算法结果（NMI）

Table 6 Result of algorithms on data set 2 (NMI)

模型	LDA	Kmeans++	GNMF	TNMF	GSDMM	BTM
Title	0.259 8 (±0.000 2)	0.171 9 (±0.000 5)	0.194 8 (±0.000 1)	0.217 3 (±0.000 4)	0.367 0 (±0.000 3)	0.362 9 (±0.000 1)
Co-query enhancement	0.298 7 (±0.000 3)	0.234 5 (±0.000 3)	0.265 9 (±0.000 1)	0.220 8 (±0.000 7)	0.309 8 (±0.000 5)	0.353 3 (±0.000 6)
Co-click enhancement	0.372 5 (±0.000 2)	0.308 8 (±0.000 5)	0.314 8 (±0.000 1)	0.176 4 (±0.000 6)	0.398 5 (±0.000 8)	0.393 6 (±0.000 2)
All enhancement	0.346 8 (±0.000 3)	0.269 7 (±0.000 1)	0.312 9 (±0.000 3)	0.253 8 (±0.000 5)	0.332 1 (±0.000 5)	0.369 1 (±0.000 2)

4 结束语

随着网络信息的爆炸式增长，文本聚类算法在很多数据挖掘工作中都发挥着越来越重要的作用，比如话题发现、个性化推荐、有效检索等。本文提出了利用网络视频的多源数据构建文本扩展模型，从视频标题、共点击视频、相关查询词等多角度进行补充表示，最后应用文本聚类算法在对文本进行划分的同时实现网络视频的聚类。在两个数据集上的多个实验验证了本文方法的有效性，进一步印证了利用外部信息进行扩展可以一定程度地提高网络视频聚类性能。但短文本聚类方法在本文提出的模型上效果欠佳，将在今后工作中进一步研究以提高改进。同时本文局限在利用文本信息扩展进行视频表示，在之后的研究工作中考虑将共点击视频的网络结构和视频内容相结合，展开更深层次的研究。

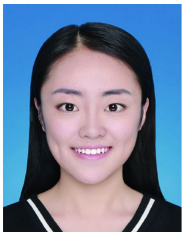
参考文献：

[1] WU X, ZHAO W L, NGO C W. Towards google challenge: combining contextual and social information for web video categorization[C]//International Conference on Multimedia 2009. Vancouver, Canada, 2009: 1109–1110.

[2] YANG L, LIU J, YANG X, et al. Multi-modality web video categorization[C]//ACM Sigmm International Workshop on Multimedia Information Retrieval. Augsburg, Germany,

2007: 265–274.
[3] HINDLE A, SHAO J, LIN D, et al. Clustering Web video search results based on integration of multiple features[J]. World wide web, 2011, 14(1): 53–73.
[4] NGUYEN P Q, NGUYEN-THI A T, NGO T D, et al. Using textual semantic similarity to improve clustering quality of web video search results[C]//2015 IEEE Seventh International Conference on Knowledge and Systems Engineering (KSE). Ho Chi Minh, Vietnam, 2015: 156–161.
[5] LIU S, ZHU M, ZHENG Q. Mining similarities for clustering web video clips[C]//International Conference on Computer Science and Software Engineering. Wuhan, China, 2008: 759–762.
[6] KAMIE M, HASHIMOTO T, KITAGAWA H. Effective web video clustering using playlist information[C]// Proceedings of the 27th Annual ACM Symposium on Applied Computing. Trento, Italy, 2012: 949–956.
[7] HUANG H, LU Y, ZHANG F, et al. A multi-modal clustering method for web videos[J]. Communications in computer and information science, 2013, 320: 163–169.
[8] ZHANG D Q, LIN C Y, CHANG S F, et al. Semantic video clustering across sources using bipartite spectral clustering [C]//IEEE International Conference on Multimedia and Expo. Taipei, China, 2004: 117–120.
[9] ZHANG J R, SONG Y, LEUNG T. Improving video classification via youtube video co-watch data[C]//Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access. Scottsdale, USA, 2011: 21–26.
[10] YIN J, WANG J. A dirichlet multinomial mixture model-based approach for short text clustering[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 233–242.
[11] YAN X, GUO J, LIU S, et al. Learning topics in short texts by non-negative matrix factorization on term correlation matrix[C]//Proceedings of the 2013 SIAM International Conference on Data Mining. Austin, USA, 2013: 749–757.
[12] SAHAMI M, HEILMAN T D. A Web-based kernel function for measuring the similarity of short text snippets[C]// International Conference on World Wide Web, WWW 2006. Edinburgh, Scotland, UK, 2006: 377–386.
[13] YIH W, MEEK C. Improving similarity measures for short segments of text[J]. Proceedings of artificial intelligence, Pune, India, 2007: 1489–1494.
[14] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering short texts using wikipedia[C]//SIGIR 2007: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, the Netherlands, 2007: 787–788.
[15] GABRILOVICH E, MARKOVITCH S. Feature generation for text categorization using world knowledge[C]// International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc, 2005: 1048–1053.
[16] HU X, SUN N, ZHANG C, et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge[C]//ACM Conference on Information and Knowledge Management 2009. Hong Kong, China, 2009: 919–928.
[17] HOTHO A, STAAB S, STUMME G. Wordnet improves text document clustering[C]//Proceedings of Semantic Web Workshop, the 26th annual International ACM SIGIR Conference. Toronto, Canada, 2003: 541–544.
[18] SONG Y, WANG H, WANG Z, et al. Short text conceptualization using a probabilistic knowledgebase[C]// Proceedings of the, International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 2330–2336.
[19] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993–1022.
[20] YANG L, QIU M, GOTTIPATI S, et al. CQArank: jointly model topics and expertise in community question answering[C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013:99–108.
[21] ARTHUR D, VASSILVITSKII S. k-means++:the advantages of careful seeding[C]//Eighteenth Acm-Siam Symposium on Discrete Algorithms 2007. New Orleans, USA, 2007: 1027–1035.
[22] CAI D, HE X, HAN J, et al. Graph regularized nonnegative matrix factorization for data representation[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(8): 1548–1560.
[23] YAN X, GUO J, LAN Y, et al. A biterm topic model for short texts[C]//International Conference on World Wide Web. Rio, Brazil, 2013: 1445–1456.

作者简介:



刘璐, 女, 1994 年生, 硕士研究生, 主要研究方向为数据挖掘、文本聚类。

贾彩燕, 女, 1976 年生, 教授, 博士生导师, 博士, 中国人工智能学会“粗糙集与软计算专业委员会”委员, 主要研究方向为数据挖掘、社会计算、生物信息学。发表学术论文 50 余篇。