基于关联规则挖掘的跨网络知识关联及协同应用

黄晓雯[1] 严 明[1] 桑基韬[1] 徐常胜[1,2]
(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)[1]
(中国-新加坡数字媒体研究院 新加坡 119615)[2]

摘 要 随着社交媒体的兴起,各种社交媒体服务应运而生,社交媒体多源化现象越来越明显。一种基于关联规则挖掘的方法可以用来分析研究社交媒体多源现象,即通过同一个用户与不同社交媒体上多源数据的行为交互,挖掘社交媒体多源数据知识关联,进而设计跨网络协同的视频推荐应用。本研究框架主要分为3个步骤:(1)基于主题建模的知识发现,对用户和视频进行主题建模,得到其在主题层上的表示;(2)基于关联规则挖掘的跨网络知识关联,以跨网络共同用户作为连接不同网络的桥梁,利用关联规则的方法挖掘不同网络间的知识关联;(3)基于跨网络知识发现的冷启动视频推荐,将用户和视频映射到同一主题空间并进行主题匹配,最终进行视频推荐。实验结果表明,通过跨网络用户协同,该跨网络知识关联方法能得到除了语义关联外更加灵活有效的跨网络关联,并在冷启动的跨网络视频推荐中取得较好的推荐效果。

关键词 跨网络关联,关联规则挖掘,视频推荐
中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.008

Association Rules Mining Based Cross-network Knowledge Association and Collaborative Applications

HUANG Xiao-wen[1] YAN Ming[1] SANG Ji-tao[1] XU Chang-sheng[1,2]
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[1]
(China-Singapore Institute of Digital Media, Singapore 119615, Singapore)[2]

Abstract Nowadays, with the rise of social media, various and disparate social media services spring up like mushrooms. As a result, the social media variety phenomenon has become more and more pervasive. In this paper, we proposed a novel association rule-based method to investigate into this social media variety phenomenon, which aims to mine the cross-network knowledge association by leveraging the collective intelligence of plenty of cross-network overlapped users. A cold-start video recommendation application was further designed based on the derived cross-network knowledge association. Three stages are mainly involved in the framework: (1) heterogeneous topic modeling, where YouTube videos and Twitter users are modeled in topic level; (2) association rule-based knowledge association, where overlapped users serve as bridge between different social media networks and a novel association rule-based method is used to derive the topic correlation between different networks; (3) cold-start video recommendation, where the Twitter users and YouTube videos are transferred to the same topic space and matched on topic level. The experiments on a real-world dataset demonstrate the effectiveness of the proposed association method, which is able to capture some more flexible knowledge association beyond the semantic association. Moreover, the performance of the cold-start video recommendation application is also very promising.

Keywords Cross-network association, Association rules mining, Video recommendation

1 引言

随着社交媒体的兴起,各种社交媒体服务应运而生,同时也产生了各种各样的社交媒体数据,比如微博类网站的文本信息流数据、媒体分享网站的多媒体数据、社交网站的用户交互数据、签到网站的地理位置数据、购物网站的消费数据等。这些社交媒体多源数据从不同角度记录着人们的网络生活,并映射着物理世界。理解社交媒体多源现象对于社交媒体分析和应用具有重要意义[1]。

社交媒体多源数据来源于人们的日常网络行为信息,在社交媒体背景下人们的行为往往分布在多个分散的网络平台。为了更好地研究社交媒体多源现象,全面地理解用户行

到稿日期:2015-11-15 返修日期:2016-01-03
黄晓雯(1993-),女,硕士,主要研究领域为社会媒体分析、多媒体检索和数据挖掘,E-mail:xiaowen.huang77@gmail.com;严 明(1989-),男,博士,主要研究领域为社会媒体分析、多媒体检索和数据挖掘,E-mail:ym646016681@gmail.com;桑基韬(1985-),男,博士,助理教授,主要研究领域为社会媒体分析、多媒体检索和数据挖掘,E-mail:jtsang@nlpr.ia.ac.cn;徐常胜(1969-),博士,教授,博士生导师,主要研究领域为多媒体分析/索引/检索、模式识别和计算机视觉,E-mail:csxu@nlpr.ia.ac.cn。

为，需要对不同网络的知识进行跨网络关联，使不同平台的信息能够互相辅助和协同利用。现有的工作从不同角度挖掘社交媒体多源数据关联：荷兰代尔夫特理工大学的学者[3]分析了用户在 Flickr、Twitter 和 Delicious 不同平台上使用标签的差异性和重叠率，并通过聚合不同平台上同一个用户的标签信息来进行跨网络推荐；Fabian Abel 等人[4]考虑用相关的新闻语料对 Twitter 用户稀疏的推文信息进行扩充，并分析比较了用实体、主题和标签分别对用户建模的语义相关性；英国南安普敦大学的学者[5]分析了用户在 Flickr 和 Delicious 上较低的标签使用重叠率上，进一步提出多种方法对不同平台用户的标签信息进行校正以进行跨平台用户标签聚合；Suman Deb Roy 等[2]利用 Twitter 社会媒体信息流的实时性和社交性来辅助进行 YouTube 上各种视频应用的设计。现有的工作大多通过语义相关性进行多平台的关联，如上文所述标签信息的聚合，但单纯的语义关联不够灵活，无法捕捉到社交媒体多源数据情景下复杂的关联模式，如用户属性关联和地理位置关联等。同时，社交媒体数据往往是异构的、冗余的，不同网络上用户行为的侧重点和表现形式不同：可能是文本、音频、图像或视频，可能是注册信息、社会网络关系或分享、上传、评论、标注等行为，单纯从语义出发无法将不同网络异构的数据知识源进行有效的关联。

为了解决这个问题，提出一种全新的利用跨网络关联用户（即在不同网络均有平台账号的用户）的集体智慧来挖掘多源数据关联的方法。为了享受不同网络的社交媒体服务，一个人往往会同时参与到不同网站中。比如，在人人网与朋友保持联系、进行交流，在新浪微博关注和跟踪热点事件，在豆瓣分享喜欢的音乐、电影、图书，在大众点评分享喜欢的餐馆和美食的图片和评论等。同一个用户在不同网络的行为信息为分析和利用社交媒体多源数据提供了数据支持。比如，分析街旁网签到行为和豆瓣网图书音乐电影关注行为可以了解不同地域人们的兴趣图谱，分析微博行为和购买行为的关联可以在新浪微博定向投放淘宝广告等。共同用户在不同网络的行为信息是社交媒体多源数据的来源，也是分析应用社交媒体多源数据的目标主体，挖掘社交媒体多源数据的关联可以更好地深入理解用户行为。本文提出一种基于关联规则挖掘的方法来分析研究社交媒体多源现象，通过同一个用户与不同社交媒体上多源数据的行为交互，挖掘社交媒体多源数据知识关联，为协同分析和应用提供解决方案。本文通过关联规则建立了跨网络知识关联，挖掘用户在不同社交网络的行为信息关联，使不同网络的用户信息和兴趣分布可以互相转换，并基于此设计了一种冷启动的跨网络视频推荐应用，实验结果表明相较于纯基于语义的方法，基于这种关联规则的方法的结果有较好的提升。

利用跨网络共同用户挖掘不同社交媒体网络的知识关联，主要的研究难点有如下两方面：(1)社会媒体平台之间的用户关联是隐式的。用户在不同平台的账号大多是不一样的，由于缺乏功能上的支持，难以获得不同网络共同用户的相关信息，给数据的采集带来极大的不便。(2)社会媒体不同平台的用户信息是异构、冗余的。此外，这些用户信息可能是冗余甚至是彼此矛盾的，如何融合不同平台的用户信息是本文的难点之一。本文采用主题模型将用户在不同社交媒体平台的行为信息转换到主题层面，再通过挖掘主题间的关联规则来融合不同平台的异构知识。

最后，将本文的贡献归纳为以下几个方面：(1)本文提出的构建跨网路关联知识的方法利用跨网络关联用户的集体智慧，使不同网络的异构行为能在用户层上进行跨网络关联，同时通过引入主题模型和用户感知，使该关联突破语义关联的局限性，在更细的粒度下进行感知。相比于单纯的语义关联等方式，本文的关联模式更加灵活，跨网络用户信息包含语义关联、地理位置关联、人物属性关联等，能更大程度上准确预测用户偏好。(2)本文基于关联规则挖掘社交媒体多源数据知识关联，解决多平台信息融合的问题，为跨网络推荐提供了可能性。(3)本文根据挖掘出的跨网络关联模式设计了一种跨网络的冷启动视频推荐应用，有效解决了用户在 YouTube 上视频推荐的冷启动问题。

2 基于关联规则挖掘的跨网络知识关联

跨网络知识关联旨在将不同平台的异构知识通过某种模式相整合，将不同网络平台互相关联，为跨平台分析和应用、满足用户需求服务。本文利用跨网络的共同用户在不同社交媒体平台的行为信息，采用关联规则挖掘不同平台的关联模式。用户的行为信息是社交媒体多源数据的来源，通过分析跨网络共同用户在不同社交媒体的行为信息，我们能在更大程度上整合社交媒体的多源数据，发掘不同平台上异构知识的关联。本文采用视频分享网站 YouTube 和社交网站 Twitter 两个平台进行相关分析。为了挖掘 YouTube 平台和 Twitter 平台的知识关联，需要了解用户在这两个平台上的行为信息，本文利用主题模型构建用户在相应平台的主题分布。得到用户的主题分布后，利用关联规则挖掘两个平台的关联模式。本文算法具体描述如下：

(1)基于主题建模的知识发现。利用主题模型生成用户在 YouTube 和 Twitter 主题层上的表示。

(2)基于关联规则的跨网络关联模式构建。选取 YouTube 和 Twitter 的跨网络关联用户，提取共同用户的主题分布信息，通过关联规则挖掘跨网络关联矩阵 T，构建跨网络关联模式。

假定有一个 YouTube 视频集合 V，对于每个 $v\in V$，可以将其表示为一个二元组(w_v, f_v)；有 Twitter 用户的推文，对于每一个 Twitter 用户，可以表示为其推文中名词和标签单词的集合。基于关联规则挖掘跨网络知识关联的目标是：给定 Twitter 平台的用户行为信息及 YouTube 平台的视频信息，能得到 Twitter 和 YouTube 的跨网络关联模式 T。

2.1 基于主题模型的知识发现

不同网络上信息往往是异构的，为了从这些异构的信息中提取有用知识，并进行知识关联，首先需要对各个网络上的知识进行有效的表示和建模。本文采用生成式的主题模型分别对 YouTube 视频和 Twitter 网络用户进行主题建模，分别得到 YouTube 视频和 Twitter 网络用户在各自主题空间的泛化表示[6]。

2.1.1 YouTube 视频主题建模

在 YouTube 上，我们希望得到的视频主题空间能同时涵盖文本和视觉语义信息，因此采用了一种多模态的主题模型 iCorrLDA[6]对 YouTube 视频进行主题建模。具体地说，对每个 YouTube 视频 v，它都可以表示为一个二元组(w_v, f_v)，

其中 $w_v=\{w_1,\cdots,w_M\}$ 为该视频的 M 个文本单词集合，$f_v=\{f_1,\cdots,f_N\}$ 为该视频所有关键帧的视觉特征向量集合，对所有的视频进行上述二元组表示。然后利用主题模型 iCorr-LDA 对其进行多模态的主题建模。最后，每个视频 v 可以表示为其主题分布形式 $v=\{v_1,\cdots,v_{K^Y}\}$，其中 K^Y 为得到的 YouTube 视频主题空间的主题个数，$v_k=p(z_k^Y|v)$ 为视频 v 在第 k 个主题上的分布概率。

2.1.2 Twitter 用户主题建模

在 Twitter 上，用户发布的大量推文信息一定程度上反映了用户的个人兴趣和偏好，因此针对用户的推文信息进行主题建模。具体地说，将每个 Twitter 用户所发的推文汇总起来，并只保留其中的名词和标签单词，生成该用户的推文单词集合。然后，利用标准的 LDA 主题模型[7]对所有 Twitter 用户的推文单词集合进行主题建模，得到每个 Twitter 用户的主题分布表示。最终，每个 Twitter 用户 u 可以表示为 $u^T=\{u_1^T,\cdots,u_{K^T}^T\}$，其中 K^T 为得到的 Twitter 用户推文主题空间的主题个数，$u_k^T=p(z_k^T|u)$ 是用户 u 在第 k 个主题上的分布概率。

2.2 基于关联规则的跨网络关联模式构建

由于不同网络的行为和信息都是由用户创造的，同一个用户在不同网络的行为应该具有一定的关联性，因此我们希望通过以同一个用户在视频分享网络和社交网络上主题分布的一一对应关系为约束，通过大量跨网络关联用户的集体智慧来求得跨网络主题间的关系。

为此，首先需要得到 YouTube 用户的视频主题分布，YouTube 用户的主题分布可以通过聚合相应用户所感兴趣的视频的主题分布得到。具体地说，对 YouTube 用户 u，将其上传的视频、点赞的视频以及播放列表中的视频聚集起来，作为其感兴趣的视频集 $v_u\in v$。已知 YouTube 视频 $v\in v_u$ 和该视频的主题分布 $p(z_k^Y|v)$，通过简单的推导，可以得到用户 u 的主题分布如下：

$$p(z_k^Y|u)=\sum_{v\in v_u}\frac{N_v(f)+N_v(w)}{N(f)+N(w)}\cdot p(z_k^Y|v) \tag{1}$$

其中，$N_v(f)$，$N_v(w)$ 为视频 v 的关键帧特征向量和文本单词的总数量，$N(f)=\sum_{v\in v_u}N_v(f)$，$N(w)=\sum_{v\in v_u}N_v(w)$ 分别表示视频集 v_u 里的所有关键帧特征向量和文本单词的总数量。通过用户主题分布聚和，最终得到所有 YouTube 用户的主题分布矩阵 $U^Y=\{u_1^Y,\cdots,u_{|u^Y|}^Y\}$。

由上文可知用户在 Twitter 及 YouTube 上的主题分布。我们发现将信息论中的互信息量作为指标能更准确地度量两个对象之间的相互性。在概率论和信息论中，两个随机变量的互信息(Mutual Information，MI)或转移信息是变量间相互依赖性的量度。Yongzheng Zhang 等人利用互信息量计算用户在 Facebook 平台上喜欢的品牌和在 eBay 上购买的品牌的相关性[7]。本文借鉴该方法，采用 PMI[1)] (Pointwise Mutual Information，点间互信息量)来计算跨网络主题关联。其定义如下：

$$pmi(z_i^T;z_j^Y)=\log\frac{p(z_i^T,z_j^Y)}{p(z_i^T)\cdot p(z_j^Y)} \tag{2}$$

其中，$p(z_i^T,z_j^Y)$ 是主题 Twitter 主题 i 和 YouTube 主题 j 的联合概率分布函数，$p(z_i^T)$ 和 $p(z_j^Y)$ 分别是主题 i 和 j 的边缘分布概率函数。

假设 $U^{T,Y}=U^T\cup U^Y$ 是 Twitter 和 YouTube 上的共同用户。通过累积所有跨网络共同用户 $U^{T,Y}$ 在 Twitter 和 YouTube 网络上的主题分布，可以得到第 i 个 Twitter 主题 z_i^T 和第 j 个 YouTube 主题 z_j^Y 间的关联程度如下：

$$\begin{aligned}&pmi(z_i^T;z_j^Y)\\&=\log\frac{\sum_{u\in U^{T,Y}}[p(z_i^T|u)\cdot p(z_j^Y|u)\cdot p(u)]}{\sum_{u\in U^{T,Y}}[p(z_i^T|u)\cdot p(u)]\cdot\sum_{u\in U^{T,Y}}[p(z_j^Y|u)\cdot p(u)]}\\&=\log\frac{|U^{T,Y}|\cdot\sum_{u\in U^{T,Y}}[p(z_i^T|u)\cdot p(z_j^Y|u)]\cdot p(u)]}{\sum_{u\in U^{T,Y}}[p(z_i^T|u)]\cdot\sum_{u\in U^{T,Y}}[p(z_j^Y|u)]}\end{aligned} \tag{3}$$

其中，$p(u)$ 为用户先验分布，这里假设其服从均匀分布，即 $p(u)=\frac{1}{U^{T,Y}}$；$|U^{T,Y}|$ 表示 Twitter 和 YouTube 上的共同用户数；$p(z_i^T|u)$ 和 $p(z_j^Y|u)$ 为上述主题建模部分输出的用户概率分布。

通过计算所有跨网络主题对间的互信息量，可以得到 Twitter 主题与 YouTube 主题的关联模式：

$$T=\{pmi(z_i^T;z_j^Y)\}_{K^T\times K^Y} \tag{4}$$

将其称为关联矩阵。给定一个新用户 u_{new} 及其在 Twitter 上主题 i 的分布 $p(z_i^T|u_{new})$，通过关联矩阵即可得到该用户在 YouTube 上主题 j 的分布为：

$$p(z_j^Y|u_{new})=\sum_{i=1}^{K^T}pmi(z_i^T;z_j^Y)\cdot p(z_i^T|u_{new}) \tag{5}$$

由此，完成了两个不同的社交媒体网络的知识关联。有了这种跨网络的知识关联，就可以由用户的 Twitter 推文分布推断出用户在 YouTube 平台上的视频偏好，了解用户在不同平台的异构行为信息，为更好地服务用户创造条件。

3 基于跨网络知识发现的冷启动视频推荐

通过上文可知，这种跨网络关联模式可以将用户的行为分布进行转移。未知用户在目标领域行为分布的情况下，通过跨网络知识关联，利用用户在已知领域的行为分布，可以预测用户在目标领域的偏好。为了验证该跨网络关联模式的有效性，设计了以下所述基于该关联模式的跨网络冷启动视频推荐应用。跨网络推荐是推荐系统在跨网络平台上的应用，跨网络的推荐系统有非常重要的现实意义，比如，分析微博行为和购买行为的关联可以使商家在新浪微博定向投放淘宝广告等。目前已有许多跨网络推荐的方法：P. Winoto 和 T. Tang 将某领域的数据集应用在 KNN 模型中，对用户在目标领域进行评分预测[9]；Pan 等利用 TCF 将二进制形式的辅助数据转换为目标数值评分矩阵[10]。

本文则利用上文所述跨网络关联模式对 Twitter 用户在冷启动情况下推荐 YouTube 视频。冷启动推荐是跨网络推荐中的难题之一，因为冷启动情况下我们无法获知用户的兴趣偏好，大大增加了跨网络推荐的难度。但在跨网络知识关联的基础上，不同网络的知识可以进行相互转化和利用，基于此关联模式，冷启动推荐应用成为可能。

3.1 用户主题分布跨网络转移

通过上文所述方法，得到 Twitter 和 YouTube 平台的跨

1) Pointwise mutual information, http://en.wikipedia.org/wiki/Pointwise_mutual_information, 2015, 4, 24

网络知识关联模式 T。为了能给 Twitter 用户准确推荐 YouTube 视频，对待推荐用户 u_{rec} 进行主题建模，获得待推荐用户在 Twitter 上的主题分布 u_{rec}^{T}，利用关联模式 T，将待推荐用户从 Twitter 平台"转移"到 YouTube 平台：待推荐用户 u_{rec} 在 YouTube 平台上的主题分布为：

$$u_{rec}^{Y}=u_{rec}^{T}\times T \tag{6}$$

3.2 基于权重学习的视频推荐

经过 3.1 节获取了待推荐的 Twitter 用户 u_{rec} 在 YouTube 上的主题分布 u_{rec}^{Y}。对给定的 YouTube 候选视频集 V_{can} 及待推荐用户 u_{rec}，为待推荐用户进行视频推荐的方法之一是直接将待推荐用户在 YouTube 上的主题分布 u_{rec}^{Y} 与候选视频集在 YouTube 上的主题分布 V_{can}^{Y} 进行相似度匹配，相似度越高表明用户对该视频的喜爱程度越大。所以选取相似性最高的 k 个视频 $V_{rec}=\{V_{rec,1},\cdots,V_{rec,k}\}$ 推荐给待推荐用户。

由于不同的 YouTube 主题对视频推荐结果预测的贡献不同，为了能更加准确地预测用户喜欢的 YouTube 视频，提出一种基于权重学习的视频推荐方法。权重学习指利用跨网络共同用户及 YouTube 候选视频集在 YouTube 平台上的主题分布进行机器学习，学习出 YouTube 平台各主题的权重分布 $w=\{w_1,w_2,\cdots,w_{K^Y}\}$。我们希望通过待推荐用户的 YouTube 主题分布与 YouTube 候选视频集的主题分布进行匹配，在匹配过程中加入权重 w，提高匹配的准确度，最后选取匹配度高的 k 个视频推荐给用户。

首先，对跨网络共同用户 $U^{T,Y}$ 的行为信息进行学习。设计了一种改进的训练模式，并采用一种基于排序学习的方法 Ranking SVM[11] 对主题进行筛选。Ranking SVM 模型的形式为：$g(*,*)=w\cdot\emptyset(*,*)$，$w$ 是模型参数，$\emptyset(*,*)$ 是机器学习的特征。利用 Ranking SVM 进行学习的目的是获得一个合适的权重 w，使参与训练的检索-文档对排序最佳。设计了一种改进的训练模式，将余弦相似度做变体，调整余弦相似度算法，使其在度量两个向量的相似度时对数值更加敏感。为了使实验结果更准确，将每个向量同时减去一个数值 i，这个数值理论上为所有向量数值和的均值，在本文方法中，这个数值采用穷举法得出，穷举的范围在主题分布均值附近。将调整后的余弦相似度算法称为调整余弦算法。

$$\cos\theta_{adj}=\frac{\sum_{n=1}^{K^Y}(x_n-i)(y_n-i)}{\sqrt{\sum_{n=1}^{K^Y}(x_n-i)^2}\cdot\sqrt{\sum_{n=1}^{K^Y}(y_n-i)^2}} \tag{7}$$

利用调整后余弦相似度的特点，定义 Ranking SVM 学习的特征为：

$$\begin{aligned}\emptyset(u^Y,V_{can}^Y)&=(u^Y-i)\odot(V_{can}^Y-i)\\&=\{(u_1^Y-i)\cdot(v_{can,1}^Y-i)\}\cdot\cdots\cdot(u_{K^Y}^Y-i)\cdot\\&\quad(v_{can,K^Y}^Y-i)\}\end{aligned} \tag{8}$$

其中，u^Y 为共同用户 u 在 YouTube 上的主题分布，$u\in U^{T,Y}$；$\odot$ 表示点乘。

Ranking SVM 学习的标签为 1 和 0，当用户 u 喜欢当前训练对中的 YouTube 视频时，给定该训练对的标签为 1，否则为 0。

利用跨网络共同用户学习得到每个主题的权重 $w=\{w_1,w_2,\cdots,w_{K^Y}\}$，将权重加入调整余弦算法中。

$$\begin{aligned}\cos\theta_{weight}&=w\odot\cos\theta_{adj}\\&=\frac{\sum_{n=1}^{K^Y}w_n(x_n-i)(y_n-i)}{\sqrt{\sum_{n=1}^{K^Y}(x_n-i)^2}\cdot\sqrt{\sum_{n=1}^{K^Y}(y_n-i)^2}}\end{aligned} \tag{9}$$

用加权重的调整余弦算法 $\cos\theta_{weight}$ 计算待推荐用户的 YouTube 主题分布，与 YouTube 候选视频集主题分布的相似度。选取相似性最高的 k 个视频 $V_{rec}=\{V_{rec,1},\cdots,V_{rec,k}\}$ 推荐给待推荐用户。

4 实验

4.1 数据集

由于没有可以直接利用的跨网络数据资源，必须构建一个新的连接 Twitter 和 YouTube 的数据集。Google+的许多用户提供了他们其他社交媒体的账号链接，所以爬取了 143295 个 Google+用户，其中有 38540 个用户提供了 YouTube 账号，39400 个用户提供了 Twitter 账号，11850 个用户提供了 YouTube 和 Twitter 账号。对每一个 YouTube 用户，通过 YouTube 的 API 爬取了用户上传的视频、点赞的视频、视频播放列表及视频信息。对每一个 Twitter 用户，通过 Twitter 的 API 下载用户最近的 1000 条推文。最终的数据集包含了 11850 个跨网络共同用户，1097982 个 YouTube 视频，9253729 条推文。

4.2 实验结果与分析

4.2.1 异构主题建模

爬取大量的社交媒体数据后，通过上文所述主题模型对不同平台的用户、视频进行主题建模，得到用户行为在主题层上的表示。按照对结果期望的经验估计，选取 60 和 40 分别作为 Twitter 和 YouTube 平台上主题建模时的主题数量。为了使读者更好地理解用户行为在主题层上的表示，将部分 Twitter 和 YouTube 主题中关键词概率 $p(z_k^T|word)$ 或 $p(z_k^Y|word)$ 最大的 5 个关键词进行可视化展示。表 1 和表 2 分别展示了 Twitter 和 YouTube 平台上的 3 组主题。

表 1 3 组 Twitter 主题

主题序号	主题的前 5 个关键词(根据 $p(z_k^T\|word)$ 排序)
#50	Social marketing business brand strategy
#44	Photography camera image canon flickr
#43	Followers insight person community follower

表 2 3 组 YouTube 主题

主题序号	主题的前 5 个关键词(根据 $p(z_k^Y\|word)$ 排序)
#28	Online marketing facebook twitter website
#2	Photography digital studio light recording
#14	Car racing road drive speed motorcycle

4.2.2 挖掘跨网络知识关联模式

由上文可知用户和视频在相应平台上的主题分布。从数据集中选取 5000 个在 Twitter 和 YouTube 上都具有丰富行为信息的共同用户进行跨网络关联模式的计算。根据式(3)、式(4)，利用 5000 个跨网络共同用户在 Twitter 和 YouTube 上的行为信息，得到用户从 Twitter 主题到 YouTube 主题的跨网络关联矩阵 T。

由表 1、表 2 可以看出，跨网络关联主题中，有基于语义关联的，如主题#50和#28(关联系数为 1.2591)，主题#44

和#2(关联系数为 1.1748)；也有基于属性关联的，如主题#43和#14(关联系数为 0.9556)：由关键词可以看出，主题#43是与人物相关的主题，而主题#14 与赛车有关，从语义上看，这两者没有关联，但实际上，主题#43 突出了“粉丝”，赛车有“车迷(粉丝)”，是一种属性关系，故而 Twitter 主题#43和 YouTube 主题#14 是一种属性层上的关联。由此可以看出，采用 PMI 挖掘关联模式不仅可以在语义层对异构知识网络平台进行关联，更可以突破语义关联的局限性，捕捉到社会媒体背景下更为复杂的关联模式，使关联模式更加灵活，并且能更准确地度量两个对象之间的相互性。

4.3 基于权重学习的冷启动 YouTube 视频推荐

4.3.1 实验设置

在冷启动视频推荐应用中，利用上述 5000 个共同用户通过 Ranking SVM 进行学习，学习后得到的权重为 w。为了减小权重差距过大造成的不良影响，将 w 按比例投影到[-1，1]区间后为其分配一个缩放比例系数 α，将 $W=1+\alpha w=\{w_1,w_2,\cdots,w_K{}^Y\}=\{1+\alpha w_1,1+\alpha w_2,\cdots,1+\alpha w_K{}^Y\}$作为新的权重用在视频推荐应用中。调整余弦相似度的数值 i 和权重比例系数α，采用穷举法得出：图 1 展示了 i 和 α 变化时 F-score 的值，将超参数 i 和 α 选定在实验结果最好的点，其中 $i=0.0187$，$\alpha=11.7$。

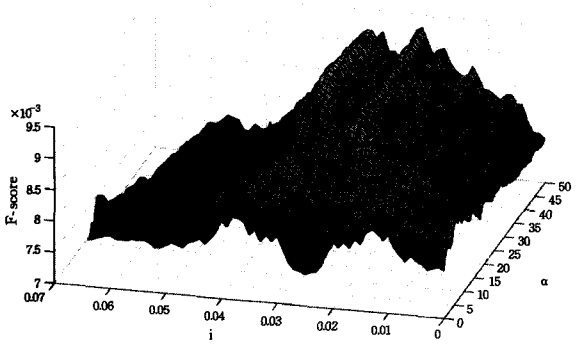

图 1 F-score 随 i，α 的变化

为了构建一个拥有足够用户视频行为的数据集用于结果评价，只保留跨网络关联用户中在 YouTube 上有超过 10 个视频行为同时在 Twitter 上发布过超过 10 个推文的用户，同时，那些不足 3 个用户关注过的视频也从视频集中移去。最终的推荐数据集中包含 2560 个待推荐用户和 4414 个候选视频，其中 2560 个待推荐用户是与 5000 个用来计算关联矩阵的用户不重合的跨平台共同用户。不重合是为了保证此视频推荐应用为冷启动视频推荐应用，待推荐用户为跨平台共同用户是为了已知待推荐用户喜欢的 YouTube 视频，并将此作为测试集，对视频推荐的准确性进行对比评价。将这些用户模拟为冷启动用户，选取待推荐用户主题和候选视频集主题相似度最高的 10 个视频推荐给用户，对推荐给用户的 10 个视频和测试集进行对比，利用信息检索的评价指标对此推荐应用进行评价。

引入信息检索的评价指标，即准确率、召回率、F-score，对此视频推荐应用进行评价[12]，通过这 3 个评价指标可以直观地判断出跨网络推荐系统的优劣。

为了评估本文基于关联规则挖掘的跨网络关联模式设计的视频推荐算法的优劣，将本文视频推荐算法与其他几种不同的算法进行对比。

- 随机推荐：随机选取 10 个视频推荐给用户。
- 基于语义匹配的推荐算法：对这 2560 个用户的推文单词与 4414 个候选视频的标签单词计算 TF-IDF，求出各待推荐用户和候选视频的 TF-IDF 后，利用调整余弦算法对它们进行相似度计算，将相似度最高的 10 个视频推荐给用户。
- 基于跨网络关联模式的直接推荐算法：待推荐用户主题和候选视频集主题直接匹配，将相似度最高的 10 个视频推荐给用户。
- 基于跨网络关联模式的加权重推荐算法：待推荐用户主题和候选视频集主题进行匹配的过程中加入机器学习得到的各主题权重$\{w_1,w_2,\cdots,w_K{}^Y\}$，并将相似度最高的 10 个视频推荐给用户。

4.3.2 实验结果

表 3 展示了上述 4 种算法的评价结果，图 2 用柱形图更加直观地展示了 4 种算法的评价指标比较。实验表明，本文所述的基于关联规则的跨网络冷启动视频推荐算法优于基于语义关联的跨网络视频推荐，准确率、召回率、F-score 均高于语义匹配算法，说明这种建立在具有集体智慧的共同用户的基础上、通过关联规则挖掘的跨网络知识关联模式能使不同网络的异构行为在用户层上进行跨网络关联，可以突破语义关联的局限性，关联模式更加灵活，并且能更准确地度量两个对象之间的相互性。该跨网络关联模式突破了不同社交媒体异构的用户信息及媒体表现形式，将不同网络平台相关联，为之后更深入了解用户行为信息及满足用户需求提供了必要的条件。此外，由实验结果可以看出，利用跨网络关联模式设计的加权重的视频推荐算法比不加权重的推荐方法结果更好，说明了主题权重优化的优势。

表 3 4 种算法的评价结果

视频推荐算法	准确率	召回率	F-score
随机推荐	0.006523	0.002215	0.003307
基于语义匹配的推荐算法	0.011602	0.004302	0.006277
基于跨网络关联模式的直接推荐算法	0.014844	0.005877	0.008420
基于跨网络关联模式的加权重推荐算法	0.016875	0.00641	0.009531

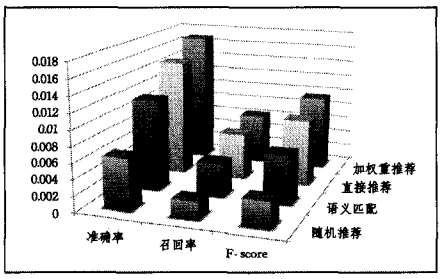

图 2 4 种推荐算法的评价指标比较

本文的跨网络知识关联模式是通过具有群体智慧的共同用户的行为信息挖掘出来的，虽然能突破单纯的语义关联，在一定程度上使用户跨网络关联更加灵活，但是在实验过程中，我们发现这种通过部分共同用户计算的关联矩阵还存在一定的随机性，与参加关联矩阵计算的用户数量及各用户行为密切相关，包括之后的基于权重的冷启动 YouTube 视频推荐，

在利用 Ranking SVM 进行学习时，如果学习的样本不够多，则可能对实验结果产生负面影响。虽然实验结果证实，通过引入这种跨网络关联模式能将两个不同平台相关联，更好地解决跨平台冷启动推荐问题，但是如何解决用户随机性造成的关联模式误差，使跨网络关联模式更加稳定有效，都还有待进一步思考和解决。在往后更深入的研究中，我们计划采用数量更加庞大的用户群进行实验，并对用户的行为信息进行筛选，选取对异构知识关联贡献更大的信息数据，以减小用户随机性造成的误差并提高实验效率，从而提高跨网络关联的准确性和有效性。

结束语 本文提出了一种基于关联规则挖掘的跨网络知识关联方法，即利用跨网络关联用户的集体智慧，运用关联规则构建异构网络平台的关联模式，使不同网络的异构行为能在用户层上进行跨网络关联，同时通过引入主题模型和用户感知，使该关联突破语义关联的局限性，在更细的粒度下进行感知。在这种关联模式下，我们提出一种基于权重学习的 YouTube 视频推荐应用，将关联模式与机器学习相结合，对新用户进行冷启动视频推荐。通过实验证明，本文提出的关联模式是有效的，这种跨网络知识关联模式能更好地理解和满足用户需求，有助于跨网络的个性化服务。在未来的研究中，我们将使用更多具有丰富行为信息的用户信息进行关联模式的构建，也期待能提出更好的算法挖掘跨网络关联模式。

参 考 文 献

[1] Sang Ji-tao, Lu Dong-yuan, Xu Chang-sheng. Overlapped user-based cross-network analysis: Exploring variety in big social media data[J]. Science Chinese, 2014, 59(36): 3354-3560(in Chinese)
桑基韬，路冬媛，徐常胜. 基于共同用户的跨网络分析：社交媒体大数据中的多源问题[J]. 中国科学，2014，59(36)：3554-3560

[2] Roy S D, Mei Tao, Zeng Wen-jun, et al. Socialtransfer: Cross-domain transfer learning from social streams for media applications [C]// Proceedings of the 20th ACM International Conference on Multimedia. Noboru Babaguchi, Kiyoharu Aizawa, 2012: 649-658

[3] Abel F, Araujo S, Gao Q, et al. Analyzing cross-system user modeling on the social Web[C]// Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. Barcelona, Spain, 2011: 28-43

[4] Abel F, Gao Qi, Houben G J, et al. Analyzing user modeling on twitter for personalized news recommendations [M] // User Modeling, Adaption and Personalization. Berlin: Springer Berlin Heidelberg, 2011: 1-12

[5] Szomszor M N, Cantador I, Alani H. Correlating user profiles from multiple folksonomies[C]// Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. Boston, Academic, 2008: 33-42

[6] Yan Ming, Sang Ji-tao, Xu Chang-sheng. Mining Cross-network Association for YouTube Video Promotion[C]// Proceedings of the ACM International Conference on Multimedia. New York, USA, 2014: 557-566

[7] Blei D M, Ng A Y, Latent M I. Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[8] Zhang Yong-zheng, Pennacchiotti M. Recommending branded products from social media[C]// Proceedings of the meeting of RecSys. 2013: 77-84

[9] Winoto P, Tang T. If You Like the Devil Wears Prada the Book, Will You also Enjoy the Devil Wears Prada the Movie? A Study of Cross-Domain Recommendations[J]. New Generation Computing, 2008, 26(3): 209-225

[10] Pan W, Liu N N, Xiang E W, et al. Transfer Learning to Predict Missing Ratings via Heterogeneous User Feedbacks[C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelli. Barcelona, Catalonia, Spain, 2011

[11] Joachims T. Optimizing search engines using clickthrough data [C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2002: 133-143

[12] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53

(上接第 18 页)

顾亚祥，丁世飞. 支持向量机研究进展[J]. 计算机科学，2011，38(2)：14-17

[33] Liang Dao-lei, Huang Guo-xing, Jin Jian. A New Multivariate Decision Tree Algorithm[J]. Computer Science, 2008, 35(1): 211-212(in Chinese)
梁道雷，黄国兴，金健. 一种多变量决策树方法研究[J]. 计算机科学，2008，35(1)：211-212

[34] Liu Jun-qiang, Sun Xiao-ying, Pan Yun-he. Survey on Association Rules Mining Technology[J]. Computer Science, 2004, 31(1): 40-47(in Chinese)
刘君强，孙晓莹，潘云鹤. 关联规则挖掘技术研究的新进展[J]. 计算机科学，2004，31(1)：40-47

[35] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425

[36] Kotsiantis S B, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques[J]. Informatica, 2009, 33(3): 249-268

[37] 国务院. 中华人民共和国计算机信息系统安全保护条例[Z]. 1994

[38] 傅建明，彭国军，张焕国. 计算机病毒分析与对抗[M]. 武汉：武汉大学出版社，2004

[39] Li Jia-jing, Liang Zhi-yin, Wei Tao, et al. A Malicious Behavior Analysis Method Based on Program Semantic[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2008, 44(4): 537-542 (in Chinese)
李佳静，梁知音，丰韬，等. 一种基于语义的恶意行为分析方法[J]. 北京大学学报：自然科学版，2008，44(4)：537-542

[40] Santos I, Brezo F, Ugarte-Pedrero X, et al. Opcode sequences as representation of executables for data-mining-based unknown malware detection[J]. Information Sciences, 2013, 231: 64-82