



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

桑基韬^①, 路冬媛^{②*}, 徐常胜^①

① 中国科学院自动化研究所, 北京 100190;

② 对外经济贸易大学, 北京 100020

* 联系人, E-mail: dongyuanlu@gmail.com

2014-04-08 收稿, 2014-07-08 接受, 2014-10-31 网络版发表

国家重点基础研究发展计划(2012CB316304)和国家自然科学基金(61303176, 61332016)资助



摘要 社交媒体大数据是大数据的重要组成部分. 与大数据的“4V”特性对应, 本文主要讨论社交媒体大数据中的 Variety-多源问题. 社交媒体的多源主要体现在不同社交媒体网络所关注的异构用户行为信息, 理解社交媒体多源现象对于社交媒体分析和社交媒体大数据的深度应用具有重要意义. 社交媒体数据具有来源于用户、服务于用户的特点. 我们提出从多个社交媒体网络的共同用户入手来进行社交媒体多源分析: (1) 跨网络用户建模, 整合分散在不同社交媒体网络的行为信息得到完整用户模型, 进行个性化服务; (2) 多源数据知识关联, 以共同用户与多源数据的交互作为桥梁, 挖掘多源数据知识关联, 服务于社交媒体协同应用.

关键词

大数据
社交媒体
多源
跨网络分析
共同用户

数据古已有之, 人类将史实刻在竹简上、图景画在瓷器上、语音录在磁带上、新闻印在报纸上. 1946年世界上第一台计算机发明后, 人类进入了信息时代, 数据有了新的载体, 其内涵也相应扩大了, 文本、语音、图像、视频等被转换为二进制形式存在于计算机中. 存储和计算领域的摩尔定律使我们能够保存并处理快速增长、体量巨大的数据. 2004年, 以Facebook为代表的社交媒体的到来引发了信息时代的新一轮革命, 人们将日常生活工作中的所感、所想都贡献到网络中, 这些社交媒体中的行为数据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%, 平均每个人所贡献的数据超过45 GB^[1], “大数据”来自社交媒体时代每个用户的“小数据”.

大数据被普遍认为具有4V(Volume-体量巨大、Velocity-增长速度、Variety-多源多样性、Veracity-真实性)特点. 在这4个特点中, Variety的研究潜力是

巨大的. 可能并不是所有人都像Facebook、沃尔玛和亚马逊等大公司数据体量、速度和真实性的体会那么深刻, 但即使是最小的问题都可以从处理多个数据源中受益. 一个数据的价值往往要通过混合另一个数据才能被挖掘并释放出来. 如将销售记录与天气、地理位置、社交媒体等数据混合后, 很容易发现影响销售收入的外在因素; 将社区房地产价格与价格历史、近期交易记录、物业动态结合, 并在社区地图上展示, 可用于预测社区房产价格走势. 此外, 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. 本文将从数据分析的角度探讨社交媒体大数据中的多源问题.

社交媒体服务的兴起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

引用格式: 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. 科学通报, 2014, 59: 3554–3560

Sang J T, Lu D Y, Xu C S. Overlapped user-based cross-network analysis: Exploring variety in big social media data (in Chinese). Chin Sci Bull (Chin Ver), 2014, 59: 3554–3560, doi: 10.1360/N972014-00292

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.对社交媒体多源现象进行研究,理解社交媒体的多源数据,是挖掘社交媒体大数据价值的重要步骤.其对于发现新规律、更全面地利用社交媒体数据、设计复杂社交媒体分析和应用等具有重要意义.比如对微博网站的信息流、社交网站的评论、媒体分享网站的图片和视频进行分析,可以从不同角度跟踪一个事件的进展和演变;对人们在视频分享网站的视频观看数据和购物网站的消费数据建立关联,有助于理解人们的兴趣和消费模型,并应用于广告定位投放.

我们提出利用跨网络的共同用户来分析社交媒体多源现象.不同的社交媒体网站关注不同的服务.为了享受各种服务,一个人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,在豆瓣网分享喜欢的音乐、电影、图书,在大众点评网分享喜欢的餐馆和美食的图片和评论等.我们称同一个人在不同社交媒体网络的用户账号为跨网络的共同用户.同一个人在不同网络的行为信息为分析和应用社交媒体多源数据提供了可能.一方面,用户是社交媒体数据分析和应用的主要对象,整合分散在不同社交媒体网络的行为信息对于深入理解用户具有重要意义;另一方面,通过同一个人与社交媒体多源数据的行为交互,我们可以挖掘社交媒体多源数据知识关联,为协同分析和应用提供了解决方案.基于跨网络的共同用户,我们在跨网络用户建模和多源数据知识关联挖掘这两方面进行了探索.

1 跨网络用户建模

一种直接利用社交媒体多源数据的方式是按照用户组织数据,即跨网络共同用户在不同网络的行为信息作为一个整体.社交媒体服务的主体是用户,趋势是个性化,用户建模是社交媒体分析的重要方面.用户在网络上的行为分散在不同网络上,要进行深入用户建模就需要整合这些行为信息.

在整合社交媒体用户的多源行为数据时,有2条思路.一个思路是针对不同的用户模型直接整合多个数据源.用户模型有很多种,如人口统计学模型、兴趣模型、社会网络模型、地理位置模型、消费模型、健康模型等.这种思路利用了数据源的互补性.以健康模型为例,我们需要收集与健康相关的多源社交

媒体信息,如 MyFitnessPal 记录了卡路里摄入、Fitbit 跟踪了每天的行走距离、Foursquare 知道去健身馆的频率、Instagram 上传了一日三餐的照片,等等.这些多源数据单独利用的价值有限,但如果通过共同用户进行组织,可以获取一个人非常完整的健康模型,进而服务于智能医疗和健康保险等方面.

另一条思路是分析不同数据源特点,利用数据源的协同性设计应用.比如我们发现,虽然 Flickr 和 Instagram 同为图片分享网站,但用户更倾向于从 Instagram 拍摄上传照片,然后转发到 Flickr 上.同样的现象出现在 Tumblr 和 Twitter 之间.我们将 Instagram, Tumblr 这种产生原始数据的社交媒体网络称为源网络,而 Flickr, Twitter 这种接收数据导入的社交媒体网络称为汇网站.掌握了这种数据源的协同关系,可以指导社交媒体数据的传播和动态演变研究,进而服务于用户行为监控和网络事件发现和跟踪等应用中.基于这种思路,我们发现了用户在 Twitter 和 YouTube 之间的行为协同性,并应用于 YouTube 的个性化视频问题中.

1.1 跨网络数据分析

我们从第三方社交媒体账号管理网站,About.me,爬取了同时提供 Twitter 和 YouTube 账号的三万用户;然后通过 Twitter 和 YouTube 提供的 API 采集了这些用户在这2个网络上的行为数据.行为数据包括用户在 Twitter 上原创和转发的推文,以及在 YouTube 上的上传、喜欢和评论的视频信息.围绕一定时间段内选定的20个社会事件,我们分析了用户在 Twitter 和 YouTube 上的反应时间先后顺序.数据分析发现,在全部20个事件上,先在 Twitter 上反应的用户数量都高于先在 YouTube 上反应的用户.图1显示了其中一

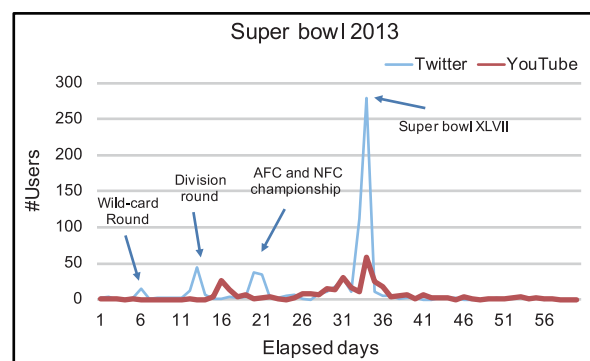


图1 跨网络数据 t 协同性分析

个事件“Super Bowl 2013”随时间变化的用户关注度曲线. 不难发现, 用户在 Twitter 网络的反应不仅时间上早于 YouTube, 且灵敏度也更高. 比如, Twitter 对 4 个子事件都有不同程度的反应, 而 YouTube 只对其中的 2 个子事件有反应, 且反应时间相对滞后.

与此同时, 数据分析还发现, 用户在 Twitter 和 YouTube 上的反应时间协同关系是类别相关的. 在科技、政治、娱乐等类别的事件上, 用户在 Twitter 的反应超前性更为显著; 而在音乐、电影等类别的事件上, Twitter 的超前性并不明显.

1.2 个性化视频推荐

在个性化应用进行用户兴趣建模时, 一般可以根据时间属性将用户兴趣分为短期兴趣和长期兴趣^[1]. 长期兴趣通常比较宽泛并相对稳定, 比如用户喜欢体育、喜欢贝克汉姆等; 而短期兴趣多由于用户自身兴趣改变或受热点事件影响而呈动态性, 比如用户由于马航失踪飞机事件开始关注大数据分析研究. 长期兴趣由于相对稳定, 可以通过分析用户注册信息、历史行为等进行挖掘; 而用户行为信息的稀疏性导致短期兴趣的转变难以捕捉. 基于以上的跨网络数据分析, 我们知道用户在 Twitter 上的行为反应要快于 YouTube, 即对于某一个特定的用户, 他/她在 YouTube 上的短期兴趣转变很可能是首先在 Twitter 上发生的. 因此我们提出利用共同用户的 Twitter 行为数据对 YouTube 上的短期兴趣进行建模. 同时, YouTube 上的行为作为长期兴趣建模的输入, 二者结合用来设计个性化视频推荐解决方案.

解决方案的框架如图 2 所示. 我们首先分析共同用户在 Twitter 的实时行为信息获得短期兴趣, 然后利用其在 YouTube 的注册信息和上传、喜欢视频历史获得长期兴趣, 最后将短期兴趣和长期兴趣结合, 通过重排序获得个性化视频推荐列表.

由于 Twitter 有 140 个字的限制, 每条推文包含的字数较少, 不同于传统主题模型假设每个文档由

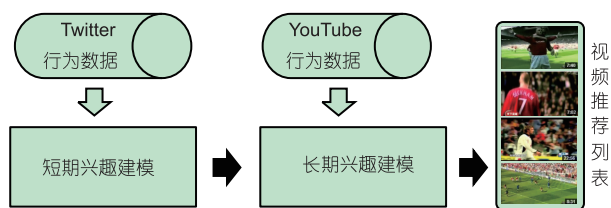


图 2 (网络版彩色)跨网络个性化视频推荐框架图

多个主题抽样产生, Twitter-LDA 假设每个文档(即推文)由一个背景模型和一个前景主题抽样得到^[2]. 在我们的短期兴趣建模问题中, 由于用户的短期兴趣很多时候与热点事件相关, 而热点事件中的关键词由于频率高很容易被误分到背景模型中. 为此, 我们先在全局时间上运行 Twitter-LDA, 得到全局的背景模型; 继而在每个局部时间段假定背景模型已知, 即已获得的全局背景模型, 重新运行 Twitter-LDA. 这样可以避免局部时间段的热点事件被作为背景模型滤掉. 经过短期兴趣建模, 我们可以获得每个共同用户在局部时间段的短期兴趣分布.

根据 Twitter 上获得的短期兴趣分布, 通过短期兴趣分布和 YouTube 视频的语义相关性计算, 我们可以为每个用户计算一个初始的视频推荐排序. 进一步, 共同用户在 YouTube 上的行为信息作为长期兴趣以重排序的方式对推荐排序进行修正. YouTube 注册信息中的关键词以及上传、喜欢视频的标题、标注、描述、类别用来表示用户的长期兴趣. 通过长期兴趣分布与 YouTube 视频匹配可以计算另一个相关性, 与之前的短期兴趣相关性线性加权, 即可获得视频的个性化相关性分数, 继而得到最终的视频推荐列表.

我们在 1000 个测试用户、277932 个 YouTube 视频上对所提出方法进行了实验评估. 用户收藏(favor)、评分(rate)、评论(comment)的视频作为 ground-truth, F-score 作为评价指标. 我们比较了以下 3 种方法: (1) TwitterTrend, 根据 Twitter 上流行的主题进行视频推荐; (2) YouTubeProfile, 根据用户的 YouTube 信息进行视频推荐; (3) T-Y Recommender, 所提出的跨网络用户建模方法进行视频推荐. 图 3 显示了实验结果. 可以看出, 通过整合从 Twitter 上分析的短期兴趣, T-Y Recommender 获得了相比其他方

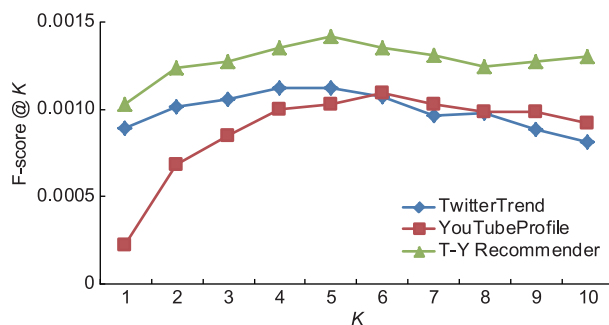


图 3 (网络版彩色)跨网络个性化视频推荐框架图

法稍好的推荐效果. 这里要注意的是, 图 3 中 3 种方法得到的 F-score 值都很低. 这是因为我们无法获得完整的用户 YouTube 行为, 如观看记录等, 导致在评估用户兴趣时的 ground-truth 很有限, 这也是个性化视频推荐评估工作共存的问题. 未来我们将研究替代信息来更好地评估所提出方法的有效性.

2 多源数据知识关联挖掘

在跨网络用户建模时, 前提是已知目标用户在不同社交媒体网络的账号对应关系, 如上述跨网络个性化视频推荐需要同时知道用户在 Twitter 以及 YouTube 的账号和对应的行为信息才能进行兴趣建模. 在实际应用时, 用户的多网络账号对应已知是个非常苛刻的假设. 而且, 不同社交媒体网络的用户不是一一对应的, 很多时候一个人并不会参与到所有的社交网络中. 基于此, 我们提出通过采集的共同用户对应关系及其在各网络的异质行为, 挖掘多源数据之间的知识层关联, 从而在应用时摆脱对显性账号对应的限制. 我们称假设共同用户账号对应的为浅层多源数据分析(如前文中基于跨网络用户建模的个性化视频推荐), 而通过共同用户账号对应挖掘得到知识关联的为深层多源数据分析.

这里所说的多源数据知识包括微博类网站的新闻关注模式、媒体分享网站的多媒体浏览模式、社交网站的社交网络模式、签到网站的地理位置模式、购物网站的消费模式等. 挖掘多源数据知识关联具有重要分析和应用价值. 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.

为更好地理解多源数据知识关联, 我们用跨媒体分析来作类比. 跨媒体分析的核心问题是寻找不同模态数据, 即图像、文本、视频、语音等之间的关联. 正如跨媒体分析和应用一样, 我们认为深度整合和应用社交媒体多源数据的首要问题也应该是多源数据的知识关联挖掘. 但与跨媒体分析不同的是, 传统多模态数据一般假设存在公共的语义子空间, 如某张照片、某段语音和某段文本描述同一个语义概念. 因此多模态的关联可以从底层进行, 通过语义特征提取或子空间映射进行跨模态分析. 然而, 社交媒体多源数据的关联模式要复杂得多, 单一的关联准则很难普遍适用, 从底层进行分析很多时候并不能

解决问题.

因此, 我们提出利用采集的共同用户账号对应关系及其在各网络的交互行为作为高层监督信息, 来挖掘多源数据的知识关联. 具体来说, 首先根据用户在每个网络的行为建立网络相关的知识空间, 然后以共同用户的账号对应作为链接, 分析异质知识空间关联. 基于此思路, 我们分析了 Twitter 大号空间和 YouTube 视频空间知识关联.

2.1 多源知识空间构建

Twitter 以其“关注-粉丝”结构和 140 个字限制成为了高效的信息传播平台. 我们称拥有大量粉丝的 Twitter 账号为 Twitter 大号. Twitter 大号由于其掌握着具有大量听众的信息发布渠道, 成为了新兴的“自媒体”. 与此同时, Twitter 大号因其对粉丝行为和观点的潜在影响成为了社交媒体背景下一种特殊的“影响力人物”. 因此, 分析 Twitter 的“关注-分析”模式具有重要意义. 在标准主题模型框架下^[3], 我们以粉丝为文档(document)、粉丝所关注的大号为词(word)进行主题建模. 得到的主题空间表达的是被粉丝共同关注的多个大号集合. 其中主题-词分布表示某个大号在一群粉丝中受欢迎的程度, 文档-主题分布表示粉丝对一个集合大号的关注程度. 表 S1 显示了所得到 Twitter 大号主题空间 80 个主题中的 3 个主题. 每个主题通过最受欢迎的 3 个大号及其用户信息来表示. 可以看出, 所发现的 Twitter 大号主题非常灵活, 既包括概括的游戏相关的主题(#43), 又包括具体的与福布斯影响人物相关的主题(#10), 还包括与地理位置相关的主题(#38).

我们拟利用 Twitter 大号的传播价值帮助 YouTube 视频的推广. 在构建 YouTube 视频主题空间时, 为同时分析视频中包含的视觉信息和文本信息(标题、标注、描述), 我们采用多模态主题模型 Corr-LDA^[20]来进行主题建模: 视频作为文档(document), 关键帧和文本元数据作为词(word). 表 S2 显示了所得到 YouTube 视频主题空间 40 个主题中的 2 个主题, 每个主题通过最相关的词和视频的标题以及关键词进行表示.

2.2 知识关联挖掘

通过知识空间构建, 我们可以获得共同用户在不同源知识空间的分布. 通过共同用户进行知识空

间关联挖掘的主要假设是：如果有大量的用户同时关注知识空间 1 的主题 A 和知识空间 2 的主题 B，那么主题 A 和主题 B 有很大可能是紧密关联的。

基于这种假设，一种直接的做法是将不同源知识空间的主题都在共同用户空间表示，然后分析多源主题在共同用户上的协同性计算主题关联。具体来说，我们将主题关联性问题建模为一个计算主题之间转移概率的问题，转移概率在所有共同用户集合上加和平均。对所有可能的多源知识空间主题对进行转移概率计算即可获得完整的多源知识关联矩阵。

另一种做法是从多源行为数据的生成原因入手。我们认为共同用户的多源行为数据实际是其用户属性在不同网络的反映，即共同用户的多源知识空间表示来源于一个共同的属性空间表示。基于此，我们仿照共轭字典学习的方法^[21]，假设存在一个独立的用户属性空间，用户属性在每个知识空间有一个表达。优化的目的是在 2 个知识空间学习这样一对共轭的字典：每个知识空间的字典能较好地重构本空间的数据，同时与另一个知识空间的字典共轭对应于统一的用户属性空间，即共同用户在共轭字典上的分布系数应该是一致的。我们设计了一种半监督的优化框架，同时利用了共同用户和各网络大量的非共同用户行为数据。

在可能的 $80 \times 40 = 3200$ 个多源主题对中，我们发现前两对最相关的主题是 {Twitter #43, YouTube #1} 和 {Twitter #38, YouTube #17}。可以看出，通过共同用户挖掘得到了复杂的多源知识关联：YouTube 的主题 #1 与游戏相关，Twitter 的主题 #43 包含了官方游戏平台 and 开发者，这一关联可认为是语义导向的；YouTube 主题 #17 和 Twitter 主题 #38 的关联则由于其关注德国，是地理位置导向的。这种复杂关联难以通过单一关联准则获得，且由于事先无法知道可能的关联包括什么，也无法通过设定多个关联准则来实现。这显示了通过共同用户来分析社交媒体多源数据的优势：完全数据导向，由用户来决定。

3 研究现状

3.1 跨网络宏观分析与应用

跨网络宏观分析不假设已知共同用户的对应关系，主要研究不同社交媒体网络的总体网络拓扑结构以及信息在相应社会网络中的传播机制。韩国先

进科技学院(KAIST)的学者分析了 Twitter 上用户之间的各种网络拓扑属性^[4]，包括用户节点度分布、最短路径、紧密性等，并从用户粉丝数、PageRank 以及转发推文数 3 方面研究用户的影响力情况。德国研究人员针对信息在社会网络中的传播范围、传播速度以及好友对信息传播的影响 3 方面分析 Flickr 的平台网络属性和图片交互情况^[5]。Ahn 等人^[6]比较了 Cyworld, Myspace 和 Orkut 这 3 个社交媒体平台的网络拓扑结构，分析了其相应的节点度分别、聚类系数、节点关联性以及网络的动态演变情况。

在信息的传播方面，南加州大学的学者先比较分析 Digg 和 Twitter 两个社交媒体平台的用户网络拓扑结构以及个体和群组的行为动态性^[7]，然后通过新闻信息在这两个网站的传播情况跟踪研究社交媒体中信息的传播机制。意大利博洛尼亚大学的研究人员在具有用户关联的 Twitter 和 FriendFeed 社会网络上分析传统的 SNA 指标^[8]，比如节点度以及最短路径等在多平台情况下的变化，用以指导跨平台用户行为分析和应用设计。

另一类宏观分析与应用的工作假设共同用户存在，但不知道显性对应关系，利用共同用户作为不同社交媒体网络的桥梁进行知识传递。香港科技大学从迁移学习的角度研究了不同网络之间用户行为的协同性，并应用于用户建模和行为预测^[9]。Zhong 等人^[10]认为不同网络之间的信息传播与共同用户有关，其在共同用户存在的假设下研究了跨网络信息传播问题。Zhong 等人^[11]假设共同用户在不同网络中存在类似的社会交互结构，进行多网络协同的关系推断研究和好友推荐应用。

3.2 跨网络微观分析与应用

与宏观分析不需要共同用户对应关系不同，跨网络微观分析需要假设已知共同用户在不同网络所对应的账号信息。第一类工作研究如何获得这种账号对应关系，称为跨网络共同用户发现。美国德克萨斯大学的研究人员首先提出多平台用户关联的问题^[12]，并提供了一个消除社会网络中隐私性和匿名性的框架，从种子用户识别到网络结构传播 2 个阶段进行关联用户发现，成功地用 Flickr 作为辅助信息源对 Twitter 的部分匿名用户进行了识别。意大利都灵大学的学者分析了多平台上不同属性对跨平台用户关联准确性的影响^[13]，提出利用多种具有辨别能

力的属性进行线性加权的方法来检测跨平台关联用户。受此工作启发, 斯坦福大学的研究人员^[14]分析了用户在不同平台 (Flickr 和 Twitter, Flickr 和 Instagram) 上注册账号名的相似性以及社会网络的重叠性, 通过综合匹配相似的账号名以及社会网络拓扑结构来进行用户关联发现。

基于共同用户对应关系, 我们可以进行跨网络微观分析并设计跨网络应用。在用户行为和社会关系分析方面, 荷兰代尔夫特理工大学的学者分析了用户在 Flickr, Twitter 和 Delicious 不同平台上使用标签的差异性和重叠率^[15], 并通过聚合不同平台上同一个用户的标签信息来进行跨平台用户建模, 以解决冷启动情况下的个性化推荐问题。英国南安普顿大学的学者发现用户在 Flickr 和 Delicious 上具有较低的标签使用重叠率^[16], 基于此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。研究结果显示, 不同平台上重叠的标签很大程度上反映了用户的一些特殊兴趣。进一步, Szomszor 等人^[17]通过引进维基百科作为外部知识源, 将用户在 Flickr 和 Delicious 上的标签文档信息进行关联分析用于用户建模。中国科学院自动化研究所在具有用户关联的 Twitter 和 Flickr 社会网络上同时

分析用户关系和行为的差异性和一致性^[18], 发现双向好友关系和共同好友行为在不同平台更容易保持一致性, 基于此设计了一种跨平台的好友推荐算法。最近, 微软亚洲研究院的研究人员与心理学家和社会学家合作, 分析了微博、豆瓣网、大众点评网上用户的行为模式, 挖掘了共同用户的行为图谱, 用以指导不同属性用户的个性化服务^[19]。

4 结论与展望

社交媒体中的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, 大量可获得共同用户对应关系为社交媒体多源分析和应用提供了宝贵数据资源。本文从跨网络用户建模和多源数据知识关联两方面讨论了对共同用户的浅层和深层利用。

社交媒体网络的共同用户可看出是一种高层监督信息。基于众包的思想, 通过整合大量共同用户在不同社交媒体网络上的行为信息进行协同分析, 可以深入理解社交媒体多源现象, 突破社交媒体跨网络分析瓶颈, 为大规模社交媒体应用提供实际解决方案。

参考文献

- 1 Xiang L, Yuan Q, Zhao S, et al. Temporal recommendation on graphs via long-and short-term preference fusion. In: Bharat R, Balaji K, Andrew T, et al., eds. Proceedings of the 16th Association For Computing Machinery SIGKDD International Conference on Knowledge Discover. Washington DC: Association For Computing Machinery, 2010. 723–732
- 2 Zhao X, Li G, Wang M, et al. Integrating rich information for video recommendation with multi-task rank aggregation. In: Sethuraman P, Balakrishnan P, Hari S, et al., eds. Proceedings of the 19th International Conference on Multimedia. Scottsdale: Association For Computing Machinery, 2011. 1521–1524
- 3 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
- 4 Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? In: Michael R, Paul J, Juliana F, et al., eds. Proceedings of the 19th International Conference on World Wide Web. Raleigh: Association For Computing Machinery, 2010. 591–600
- 5 Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network. In: Michael R, Paul J, Juliana F, et al., eds. Proceedings of the 19th International Conference on World Wide Web. Raleigh: Association For Computing Machinery, 2010. 721–730
- 6 Ahn Y Y, Han S, Kwak H, et al. Analysis of topological characteristics of huge online social networking services. In: Carey L W, Mary E Z, Peter F, et al, eds. Proceedings of the 16th International Conference on World Wide Web. Banff: Association For Computing Machinery, 2007. 835–844
- 7 Lerman K, Ghosh R. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: William W C, Samuel G, eds. Proceedings of the Fourth International Conference on Weblogs and Social Media. Washington D C: AAAI Press, 2010. 90–97
- 8 Magnani M, Rossi L. The ML-model for multi-layer social networks. In: Zhao X, Li G, eds. Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. Washington D C: IEEE Computer Society, 2011. 5–12
- 9 Zhong E, Fan W, Yang Q. User behavior learning and transfer in composite social networks. Association For Computing Machinery Transactions on Knowledge Discovery from Data (TKDD), 2014, 1: article 6

- 10 Zhong E, Fan W, Zhu Y, et al. Modeling the dynamics of composite social networks. In: Inderjit S D, Yehuda K, Rayid G, et al., eds. Proceedings of the 19th Association For Computing Machinery SIGKDD International Conference on Knowledge Discover. Chicago: Association For Computing Machinery, 2013. 937–945
- 11 Zhong E, Xiang E W, Fan W, et al. Friendship prediction in composite social networks. arXiv preprint arXiv:1402.4033
- 12 Narayanan A, Shmatikov V. De-anonymizing social networks. IEEE Symposium on Security and Privacy, 2009. 173–187
- 13 Carmagnola F, Cena F. User identification for cross-system personalisation. Inform Sci, 2009, 1: 16–32
- 14 Ding X, Zhang L, Wan Z, et al. De-anonymizing dynamic social networks. In: Proceedings of the Global Communications Conference. Houston: IEEE Computer Society, 2011. 1–6
- 15 Abel F, Araújo S, Gao Q, et al. Analyzing cross-system user modeling on the social web. In: Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. Barcelona: IEEE Computer Society, 2011. 28–43
- 16 Szomszor M N, Cantador I, Alani H. Correlating user profiles from multiple folksonomies. In: Hypertext and Hypermedia. Boston: Academic Press, 2008, 2: 33–42
- 17 Szomszor M, Alani H, Cantador I, et al. Semantic modelling of user interests based on cross-folksonomy analysis. The Semantic Web-ISWC, 2008, 1: 632–648
- 18 Yan M, Sang J, Mei T, et al. Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge. In: Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. Barcelona: IEEE Computer Society, 2011. 1–6
- 19 Yuan N J, Zhang F, Lian D, et al. We know how you live: Exploring the spectrum of urban lifestyles. In: Amr E A, Balachander K, eds. Proceedings of the 1st Conference on Online Social Networks. Chicago: Association For Computing Machinery, 2013. 3–14
- 20 Blei D M, Jordan M I. Modeling annotated data. In: Proceedings of the 26th Annual International Association For Computing Machinery SIGIR Conference on Research and Development in Information Retrieval. Toronto: Association For Computing Machinery, 2003. 127–134
- 21 Yang J, Wang Z, Lin Z, et al. Coupled dictionary training for image super-resolution. IEEE Trans Image Process, 2012, 8: 3467–3478

Overlapped user-based cross-network analysis: Exploring variety in big social media data

SANG JiTao¹, LU DongYuan² & XU ChangSheng¹

¹ Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;

² University of International Business and Economics, Beijing 100020, China

Social media contributes much to big data. Among the “4V” characteristics of big data, this article focuses on investigating the “variety” in big social media data. Social media variety mainly concerns with the heterogeneous user behaviors in different social media networks. Understanding into social media variety plays important roles in insightful social media analysis and comprehensive social media applications. Social media is typically generated from user and designed for user services. We propose to explore social media variety by investigating the overlapped users between different social media networks. Two problems are discussed: (1) cross-network user modeling, where the scattered user behaviors are integrated for complete user modeling and personalized service development; (2) heterogeneous knowledge association, where the overlapped users serve as bridge to mine the cross-network knowledge association and applied in social media collaborative applications.

big data, social media, variety, cross-network analysis, overlapped user

doi: 10.1360/N972014-00292

补充材料

表 S1 Twitter 大号主题举例

表 S2 YouTube 视频主题举例

本文以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.