基于兴趣的社交网络用户聚类及可视化

汤 颖[1] 钟南江[1] 孙康高[1] 秦大康[2] 周伟华[3]
(浙江工业大学计算机科学与技术学院 杭州 310023)[1] (南通大学理学院 南通 226019)[2]
(浙江大学管理学院 杭州 310027)[3]

摘 要 随着社交网络的流行,从各种各样的社交网络数据中提取出有效信息并进行清晰直观的可视化分析,从而为用户提供有价值的潜在知识,显得尤为重要。聚类分析是数据挖掘中的重要分析手段,传统的面向社交网络数据的用户聚类分析大都仅考虑网络的拓扑链接结构,未考虑用户的兴趣相似度。文中基于贝叶斯概率模型来计算用户兴趣相似度并进行聚类,进一步设计交互可视化方式来展示上述聚类结果。具体地,针对社交网络中的用户评分数据建立潜在语义模型来提取表示每个用户兴趣特点的特征向量;基于用户的特征向量对用户进行聚类,得到具有不同特征的人群,并通过实验和热度图选择合适的人群聚类数;最后提出了基于层次气泡图的可视化展现和分析方案,将用户、电影类型、电影等多维信息在图形中交互展示,支持用户从全局概览到局部细节的推进式探索,从多角度可视化人群特征。对豆瓣网用户和电影评分数据进行了实验和分析,结果验证了所提方法的有效性。

关键词 社交网络,聚类,数据可视化,潜在语义模型
中图法分类号 TP391 **文献标识码** A

Clustering and Visualization of Social Network Based on User Interests

TANG Ying[1] ZHONG Nan-jiang[1] SUN Kang-gao[1] QIN Da-kang[2] ZHOU Wei-hua[3]
(School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(School of Science, Nantong University, Nantong 226019, China)[2]
(School of Management, Zhejing University, Hangzhou 310027, China)[3]

Abstract With the development of social network, it becomes more and more important to extract useful information from the social network and provide valuable knowledge to users in an interactive visual interface intuitively. Clustering, as a crucial method in data mining, offers the global data analysis results. Traditional clustering methods of social network data mainly consider network topological structure. However, they haven't considered the user interests for clustering. In this paper, the users are clustered by computing user-interest similarity based on Bayesian probabilistic model, furthermore, the interactive visualization method is designed to present the user clustering results. Specifically, we computed the feature vectors representing users' interests based on latent semantic model. Then clusters with different interest characteristics were built based on these feature vectors. The suitable number of clusters are determined by heat map visualization results. Finally, we presented the interactive visualization method based on hierarchical bubble chart to support users to explore the clustering results from the global overview to local details. We performed experiments and analysis with data crawled from Douban website. The results validate the effectiveness of our method.

Keywords Social network, Clustering, Data visualization, Latent semantic model

1 引言

随着信息技术的飞速发展和互联网的普及,社交网络焕发出新的生命和活力,大量基于社交网络的服务和应用应运而生,国内以腾讯QQ、微信和豆瓣网为代表。如何建立有效的数据分析模型来提取社交网络中蕴含的潜在知识是当前数据挖掘领域研究的热点。聚类分析是数据挖掘中非常重要和基础的分析手段,它可以将网络中具有相似特征的节点聚类在一起,从而提供对社交网络全局结构的总览和概括。

传统的社交网络用户聚类方法通常基于网络的拓扑链接结构对用户进行聚类。该类方法能够将连接紧密的用户聚类在一起,但是在聚类时未过多考虑用户的兴趣相似度。在很多应用中,我们更希望能够发现兴趣相似的用户聚类,从而更好地理解社交网络中的人们形成的子群结构。

目前许多基于用户兴趣的相似度计算方法均基于用户标签[1]或者用户个人资料[2]。但是对于没有用户标签并且也不公开用户个人资料的社交网络,我们就需要寻找其他方法来计算用户的兴趣相似度。很多社交网络带有评分功能,用户

本文受国家教育部新世纪优秀人才支持计划(NCET-13-0526),国家自然科学基金(71571160),浙江省自然科学基金(LY14F020021)资助。
汤 颖(1977-),女,博士,副教授,硕士生导师,CCF会员,主要研究方向为计算机图形图像、虚拟现实和信息可视化,E-mail:ytang@zjut.edu.cn;钟南江(1991-),男,硕士生,主要研究方向为数据挖掘、信息可视化;孙康高(1992-),男,硕士生,主要研究方向为数据挖掘、信息可视化;秦大康(1977-),男,博士,副教授,主要研究方向为机器学习、动力系统、元胞自动机,E-mail:qindk@ntu.edu.cn(通信作者);周伟华(1976-),男,博士,教授,博士生导师,主要研究方向为数据分析与管理。

对他/她看过的电影、听过的音乐、去过的饭店等给出具体评分，评分高低代表用户对物品的喜好程度。若用户的评分越相似，则我们可以认为用户在此领域的兴趣也越相似。基于社交网络上的用户评分，本文提出了结合评分信息的用户兴趣相似度计算及聚类算法。潜在语义模型(Latent Semantic Model)[3]属于概率生成式模型，是一种贝叶斯概率模型，目前主要用于对物品的评分预测和推荐。本文基于用户电影评分数据，通过潜在语义模型引入潜在变量对用户进行建模，提取出用户的特征向量，从而抽象出用户在电影方面的兴趣爱好。

为了直观地展示上述基于概率模型的用户聚类结果，帮助人们更好地理解和比较不同用户聚类的兴趣特点，在上述潜在语义模型的计算结果的基础上提出面向聚类结果展示的可视化方法。一方面，采用热度图展示比较不同聚类的静态统计信息，从而更好地确定聚类参数 k 的取值；另一方面，提出基于层次气泡图的动态交互可视化方法，展示不同层次下的聚类结果，从而让用户既能看到所有聚类分布的全貌，也可探究某一聚类中具体的用户兴趣信息。

综上所述，本文针对社交网络数据提出了结合数据分析和可视化的用户聚类方法。潜在语义模型为数据可视化提供模型基础，可视化将潜在语义模型得到的结果进行直观、交互的展示。本文抓取了豆瓣网用户以及用户电影评分数据并对其进行聚类分析，实验结果表明本文方法可以有效地对豆瓣网用户进行聚类。

本文第 2 节讨论相关工作；第 3 节介绍基于潜在语义模型的用户聚类方法；第 4 节介绍对聚类结果的可视化设计；第 5 节总结全文，并对未来的研究方向提出设想。

2 相关工作

2.1 潜在语义模型

潜在语义模型是一种统计方法——概率潜在语义分析(pLSA)[4]的泛化。因为潜在语义模型引入了潜在变量的概念，这让它看上去与一些聚类算法非常相似，但实际上潜在语义模型并没有对数据进行聚类，即使是概率上的。概率潜在语义模型在很多方面与一些降维方法以及矩阵分解方法是相似的，如奇异值分解(SVD)[5]和主成分分析(PCA)[6]。

早期的潜在语义模型只有一个潜在变量，在模型的构造中也仅仅考虑到了用户偏好的因素，但随着社交网络的发展，越来越多的信息可以被利用，比如好友关系、用户行为、物品信息等。如果能将这些信息利用起来，准确率也会得到相应的提升。

文献[7]对潜在语义模型进行扩展，主要提出了一种基于社会影响力的方法来进行推荐。在潜在语义模型中对用户的偏好进行建模以预测用户的选择。该文认为，在社交网络中，用户的选择不仅仅与他自己的偏好有关，往往还会受到朋友偏好的影响。该文主张使用社会影响力以及用户偏好进行物品推荐，并且提出了一个概率生成模型——基于社会影响力的选择(SIS)，它在物品选择的过程中量化和包含了用户与朋友之间的社会影响力，以及用户偏好和物品内容。该文提出了一种新的模型参数学习算法，用来推导出两层潜在变量，包括有影响的朋友和潜在主题。

2.2 数据可视化

Card 等人对信息可视化(Information Visualization)的定义为：对抽象数据使用计算机支持的、交互的、可视化的表示形式，以增强认知能力[8]。可视化主要研究大规模的信息资源视觉呈现[18]，以及利用图形和图像的相关技术和方法将数据直观地显示，为用户提供可交互的操作等，帮助人们理解和分析数据[19]。如今，可视化技术已成为一个基本的工具，用来揭示数据之间存在的关系和背后隐匿的信息[20]。

随着可视化技术的发展，研究者基于不同需求提出了大量的网络可视化或图可视化技术，Herman 等人[9]综述了图可视化的基本方法和技术。经典的基于节点和边的可视化是图可视化的主要形式，比如 NodeXL[21]、MatrixExplorer[22]、力导引布局[23]等。具有层次特征的图可视化的典型技术包括 H-Tree、圆锥树 Cone Tree、气球图 Balloon View、放射图 Radial Graph、三维放射图 3D Radial、双曲树 Hyperbolic Tree 等。对于具有层次特征的图，空间填充法也是采用的可视化方法，例如树图技术 Treemaps[10]是一种在受限空间内展示树状数据结构的可视化方法[11]。通过将矩形不断进行细分(Slice and Dice)，可以在固定大小区域内展示多层级的数据信息，也可以比较直观地展示同层级数据之间的比较情况，但在结果中很容易出现细长的矩形，不利于辨别。为了解决这一问题，提出了 Voronoi Treemap(泰森多边形树状结构图)[12]的方法，其可以避免出现细长矩形的情况，从而达到更好的可视化效果；而且最外层的区域也不再限制为矩形，可以在任意形状内进行多层级数据的展示。马赛克图(Mosaic Display)是一种用来展示关联表(Contingency Table)的图解法[13]。马赛克图与 Treemap 的区别是：每一次将一个矩形切分成几个矩形都等价于增加一个维度的信息。它一般用于二维、三维、四维的低维数据的可视化展示。本文采用嵌套圆圈的形式来展示结果，并将这种展示形式命名为“气泡图”[14]。用一个圆圈将构成信息包含起来，这符合集合的表示形式，也能够展示数据结果的层次关系，更为用户提供了方便的交互。与传统上一旦生成就固定不变的二维表格表现方式相比，它更加灵活多变，通过缩放操作可以为用户清晰展示用户关心的数据细节，也可以进行整体上的宏观比较。

为了实现人与机器之间的相互合作，研究者们提出了可视化交互模型。交互模型定义了人与机器在协作过程中各自承担的任务，以及他们之间消息传递的规则和方式。交互模型需要对用户与机器之间的交互元素进行定义，根据用户的不同操作，机器端做出相应的反馈。交互模型定义了人机交互任务的具体实现方式和方法，为多模社交网络数据的可视分析系统的交互设计与实现提供了重要的理论支持。

Keim 等人[15]为可视分析的交互框架给出了高层的、概念化的描述，其主要内容是对人、机器两端各自承担的任务范畴进行了最佳的划分。例如，机器这一方的任务主要是数据管理、数据挖掘、统计分析、过滤、图形渲染和绘制等；人这一方的主要任务是认知、感知、信息组织与设计、推理、决策、行动等。但是这样的交互框架没有面向任务建立真正的交互模型，仅仅是对人机交互中的概念模型做出了宏观的描述。Pike 等人[16]根据多层次的任务特点，从高层到低层的映射维度建立了信息可视分析的交互模型，其中，在用户这一方定义了一系列高层目标，如浏览、比较、分析、探索、评价、吸收、理解等；除此之外，也定义了相应的低层次任务，比如排序、过滤、检索、聚类、寻找异常点等。在机器这一侧同样定义了从高层到低层的两个层次，定义了交互的可视化界面表征元素

以及交互元素，高层元素主要偏重表征内容以及交互内容，而低层元素的定义则更加偏重表征与交互的具体技术。这种交互模型对可视分析中人、机器各自的交互元素做出了较为细致的定义，但是根据具体的实际情况，需要在交互建模的过程中与不同粒度、层次以及领域相关的任务建立联系。

本文采用热度图和层次化显示的气泡图对聚类结果进行验证和可视化分析。利用热度图可以很好地验证本文提出的聚类方法的有效性，利用气泡图则便于动态层次化地探索聚类群体的内在特性。

3　基于潜在语义模型的用户聚类

涉及到的数据包括用户集 $U=\{u_1,\cdots,u_n\}$、电影集 $Y=\{y_1,\cdots,y_m\}$ 以及一系列可能的评分 v，$v\in\{1,2,3,4,5\}$。把观测到的数据写成这样的形式：(u,y,v)，其表示用户 u 看过电影 y，并且对其的评分值为 v。评分 v 表示了用户对电影的喜爱程度，其取值为一定范围内的数字。如豆瓣网中的用户对电影评分的范围为[1,5]。评分数据集可以简洁地表示为一个 $n*m$ 的矩阵 A，矩阵的每一行表示一个用户，每一列表示一部电影，如图 1 所示，其中?表示用户未看过该电影。

	A	B	C	D
1	3	?	5	?
2	?	2	4	1
3	3	1	?	?
4	2	?	4	2
5	?	4	3	5
6	4	3	5	?

图 1　评分矩阵 A

我们的目标是根据该评分矩阵为用户聚类，将电影兴趣一致的用户归为一类。聚类的关键在于提取用户的特征向量，虽然矩阵 A 的每一行都可以用来表示对应的用户向量，但是这存在两个问题：

1）由于电影部数过多导致用户向量维度过高，影响计算效率。

2）用户未看的电影的评分未知，导致向量无法直接进行计算。若只选择用户共同看过的电影作为用户特征向量，则忽略了许多其他信息。

为了解决以上两个问题，对评分数据建立潜在语义模型，通过训练得到用户特征向量。

3.1　基于潜在语义模型的用户特征向量提取

选取的用户特征向量是否合适是聚类成功与否的关键。为了建立与用户兴趣相关的向量，对用户评分数据建立潜在语义模型(LSM)，并通过该模型提取用户的特征向量。概率 $P(u,y,v)$ 表示用户 u 看过电影 y，并且评分为 v 的概率。建立潜在语义模型的关键在于为每一组“用户-电影-评分”数据引入潜在变量 z，潜在变量 z 有 k 个不同的取值。z 可以看作是用户的聚类或者群组，每个数据元组都与潜在变量 z 有关。更加直观地来说，对于每一个数据 (u,y,v)，用户 u“因为”原因 z 而选择看电影 y，并且给出的评分为 v，那么，最终的模型可以写成如下混合模型：

$$P(u,y,v)=\sum_z P(u,y,v,z)=\sum_z P(v|y,z)P(y|z)P(z|u)P(u) \tag{1}$$

其中，$P(z|u)$ 表示用户 u 属于聚类 z 的概率，$P(y|z)$ 表示用户聚类 z 看了电影 y 的概率，$P(v|y,z)$ 表示用户聚类 z 对电影 y 的评分为 v 的概率。该模型的图形化描述如图 2 所示。

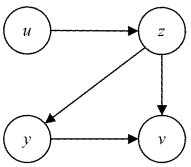


图 2　模型的图形化描述

从模型的图像化描述中可以看出，潜在变量 z 对用户看什么电影、用户对电影做出什么样的评分有着直接的影响，并且它不会忽略用户没有看过的电影，而是对它们进行了预测，我们认为它最能反映用户兴趣。因此我们可以选择用户属于每个聚类 z 的概率 $P(z_j|u)$ 作为特征值，那么用户向量为一个 k 维的向量：$(P(z_1|u),\cdots,P(z_k|u))$。

选取了用户的特征向量后，通过训练得到向量中每一维的值。本文使用最大期望估计(EM)算法对模型的参数进行训练。

EM 算法分为 E 和 M 两步。在 E 步中，计算在给定数据 (u_i,y_i,v_i) 的前提下潜在变量为 z_j 的概率分布：

$$Q(z_j;u_i,y_i,v_i)=P(z_j|u_i,y_i,v_i)=\frac{P(v_i|y_i,z_j)P(y_i|z_j)P(z_j|u_i)}{\sum_z P(v_i|y_i,z)P(y_i|z)P(z|u_i)} \tag{2}$$

在 M 步中，使用 $Q(z_j;u_i,y_i,v_i)$ 更新 E 步中的概率参数：

$$P(v_i|y_i,z_j)=\sum_{(u_i,y,v):y=y_i,v=v_i} Q(z_j;u,y,v) \tag{3}$$

$$P(y_i|z_j)=\frac{\sum_{(u,y,v):y=y_i} Q(z_j;u,y,v)}{\sum_{(u,y,v)} (z_j;u,y,v)} \tag{4}$$

$$P(z_j|u_i)=\frac{\sum_{(u,y,v):u=u_i} Q(z_j;u,y,v)}{\sum_z \sum_{(u,y,v):u=u_i} Q(z;u,y,v)} \tag{5}$$

经过 M 步的计算后再更新 E 步中的概率。通过 E 步和 M 步的不断迭代，得到所有的用户特征值 $P(z|u)$，由此得到所有用户的特征向量：$(P(z_1|u),\cdots,P(z_k|u))$。这不仅使得用户特征向量的维度得以降低，也考虑到了用户对所有电影的评分。

3.2　基于特征向量的用户聚类

3.1 小节通过建模提取出了所有用户的特征向量集 F，其中每个向量为 $(P(z_1|u),\cdots,P(z_k|u))$，其意义是用户 u 属于潜在变量 z_i 代表的聚类的概率。当潜在变量的概率个数为 5 时，就可以用一个五维向量表示该用户的特征。如用户 u_i 属于潜在变量代表的聚类 0，1，2，3 和 4 的概率分别为 0.6，0.05，0.2，0.05 和 0.1，那么用户 u_i 的特征向量表示形式为 (0.6，0.05，0.2，0.05，0.1)。将所有豆瓣用户都用这样的五维向量表示后，便可以利用 K-means 聚类算法[15]对用户进行聚类，将具有相同特征的用户聚集，形成具有特征的代表性人群。

K-means 算法属于硬聚类算法，是非常具有代表性的基于平方误差的迭代重分配聚类算法，采用距离作为节点间相似度的评价指标，认为两个对象的距离越近，那么它们的相似度就越高。K-means 算法在第一步中随机选取 k 个用户作为聚类质心点，在我们的实验中随机生成 k 个用户 ID，把这 k 个用户作为聚类质心，比如选取的聚类质心点为 $\mu_1,\mu_2,\cdots,\mu_k\in F$；接着重复以下步骤直至收敛。

(1)对于每个用户 u_i，计算其当前应归属的聚类类别：

$$c^i:=\arg\min_j \| x^{(i)}-\mu_j \|^2$$

其中，$x^{(i)}$ 表示用户 u_i 的特征向量。

(2)对于每一个聚类 j,重新计算其质心:

$$\mu_j := \frac{\sum_{i=1}^{n} l\{c^i=j\}x^{(i)}}{\sum_{i=1}^{n} l\{c^i=j\}} \tag{6}$$

通过以上的计算方法得到了用户集的 k 个聚类。

3.3 基于豆瓣网数据的聚类结果

基于豆瓣网的数据对豆瓣网络中的用户进行聚类。本文使用的数据都是基于爬虫脚本从豆瓣网抓取的真实评分数据。数据以元组的形式(用户 ID,电影 ID,对应评分)存储。在筛选去除信息不全、信息错误、评分过多或者过少的用户数据后,保留了 35911 条电影评分数据,电影评分的取值范围为[1,5],数据包含了 101 个用户和 2752 部电影。

根据用户兴趣用本节方法对用户聚类后,计算每一类用户对各种类型电影的偏爱程度来证明用户聚类的有效性。每一类用户对某一类型电影的偏爱程度可以根据用户对电影的评分来计算,例如评分数据(u_i,y_i,v_i)表示用户 u_i 对电影 y_i 的评分为 v_i,用户 u_i 属于聚类 k,电影 y_i 的类型有喜剧/爱情。由于豆瓣网络的评分范围为[1,5],因此选择 3 为基准点,若 $v_i>3$,那么聚类 k 对喜剧、爱情的偏爱程度加 1;若$v_i<3$,那么聚类 k 对喜剧、爱情的偏爱程度减 1;若 $v_i=3$,不作任何操作。使用 L_{pq} 表示用户聚类 p 对某一电影类型 q 的偏爱程度,那么 L_{pq} 的函数表达式如下:

$$L_{pq}=\sum_{u_i\in p}\sum_{y_i\in q} f(u_i,y_i,v_i) \tag{7}$$

$$f(u_i,y_i,v_i)=\begin{cases}1, & v_i>3\\ 0, & v_i=3\\ -1, & v_i<3\end{cases} \tag{8}$$

通过上式,将每个用户聚在不同电影类型上的偏爱程度的累加值除以总的评分次数做归一化,从而得到最终的聚类对各个电影类型的偏爱值,如表 1 所列。

表 1 不同用户聚类对各个电影类型的偏爱程度

电影类型	聚类 0	聚类 1	聚类 2	聚类 3
爱情	2.617	0	1.166	5.125
传记	0	3.218	0.851	1.888
动画	1.1424	0	2.818	7.718
动作	0.596	3.305	0	0.218
短片	0	0.911	2.021	4.112
儿童	0.785	0	2.752	2.667
犯罪	0	2.458	1.162	6.965
歌舞	0.842	0	1.767	3.260
古装	1.603	2.454	0.592	0
家庭	0	0.954	2.616	4.645
惊悚	1.840	0	0.811	4.919
科幻	1.716	0.854	0	
功夫	1.562	2.714	0	0.761
历史	1.455	4.245	0	0.688
冒险	1.941	0	0.780	5.345
奇幻	2.071	0	0.890	4.783
情色	1.590	2.388	0	0.790
同性	1.012	0	2.060	3.300
武侠	1.389	2.454	0.690	0
西部	0	0.742	1.550	2.460
喜剧	2.193	0	0.900	7.673
悬疑	0	0.877	2.020	6.910
音乐	1.903	0	0.907	3.546
运动	0.765	0	1.540	2.600
灾难	2.230	1.411	0	0.700
战争	0	2.270	0.960	3.680

从表 1 中可以看出,不同的用户聚类有着不同的喜好,但是表格的形式并不直观,不利于用户理解每个聚类兴趣的特点。下面给出可视化界面设计,以帮助用户更好地理解聚类计算结果。

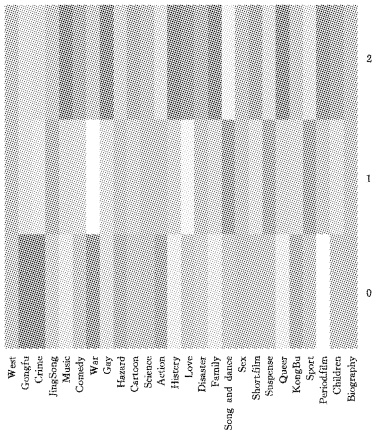
4 用户聚类结果的可视化

本节将通过可视化来展示聚类结果,其中包括静态的可视化结果和基于层次气泡图的动态可视化。在静态的可视化中,基于热度图展示各个用户聚类对不同电影类型的偏爱程度;在动态的可视化中,基于层次气泡图展示各个用户聚类喜欢的电影类型和电影,并设计交互以供用户自主探索。

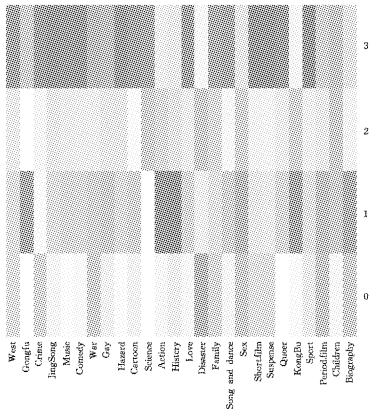
4.1 静态的统计结果可视化

表 1 展示了当聚类个数为 4 时不同用户聚类对不同电影类型的偏爱程度,但表格的形式并不能直观地显示出某种规律,我们可以借助热度图的形式展示各个聚类对不同电影类型的偏爱程度,以证明聚类的有效性。

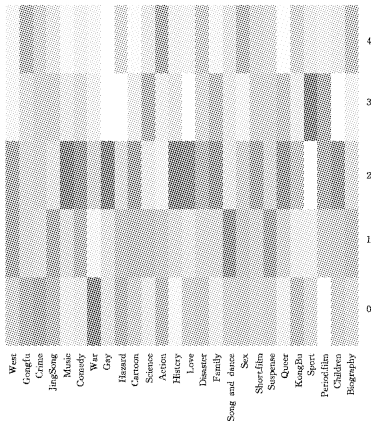
通过热度图的方式对聚类结果进行可视化的展示,如图 3 所示。



(a)用户聚类个数 $k=3$



(b)用户聚类个数 $k=4$



(c)用户聚类个数 $k=5$

图 3 不同用户人群对电影的偏好

图 3 通过热度图的形式分别展示了用户聚类个数为 3，4 和 5 时不同用户聚类对各个电影类型的偏爱程度，行表示用户聚类人群，列表示 26 类电影，每个色块表示该人群对某一类型电影的偏爱程度。色块的颜色表示的意义如图 4 所示，色调越暖表示用户越喜爱这种类型的电影；色调越冷表示用户越不喜欢这种类型的颜色。若颜色接近白色，则表示用户对这种类型的电影没有明显的倾向。

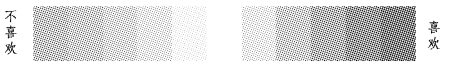


图 4　热度图色带示意图

通过热度图可视化，我们可以有效地选择聚类个数参数的值。比较图 3 中的 3 张子图可以发现，当聚类个数为 3 时，用户聚类分得不够细致；当聚类个数为 5 时，用户聚类过于细致，导致聚类 0 和聚类 4 过分拟合；而当用户聚类个数等于 4 时，我们认为这是最合适的选择。从用户聚类个数为 4 的结果(图 3(b))中可以看出，人群 0 对爱情(Love)、同性(Gay)、情色(Sex)等与爱相关的类型的电影具有明显的偏爱；而人群 1 则更加偏爱动作(Action)、功夫(Gongfu)、战争(War)、古装(Period film)等充斥着激情暴力的年代电影；人群 2 偏爱的电影类型则比较温馨和温暖，如卡通(Cartoon)、儿童(Children)、家庭(Family)等类型；人群 3 喜欢的电影类型很多，但是不包括动作、功夫、情色等暴力激情的电影类型。

通过表格和热度图的展示，我们发现该方法可以根据用户的兴趣对用户进行有效的聚类。相比于表格形式，热度图的展示让我们更加方便直观地观察到每个聚类的兴趣特征。但是这种静态的图片让用户获取到的信息有限，只能让用户对聚类结果有一个总体的概览，而不能观察到每个用户聚类的细节，比如从热度图中只能观察到用户喜欢哪种类型的电影，而不知道每个用户聚类喜欢的是哪些具体的电影。

4.2　基于层次气泡图的交互可视化

对社交网络中的数据进行有针对性的筛选和处理后，不应该仅仅是一幅由程序计算后生成的图片或表格，而最好是一个可以进行交互的应用，使得用户可以进行方便的操作，充分发挥人的认知能力。用户可以根据自己的视角来获取感兴趣的内容，通过交互的方式逐步缩小兴趣点的范围。

基于热度图的可视化可以在一定程度上展示数据分析结果。为了使得用户能够交互地、多层次地探索数据分析结果，在网页端设计了层次气泡图，如图 5 所示。层次气泡图由嵌套的圆圈组成，一共 4 层。最外层圆圈代表总体人群，中间一层的 4 个圆圈表示 4 个用户聚类，每个用户聚类中包含的圆圈表示该聚类喜欢的不同电影类型，最里面一层的 5 个小圆圈代表不同电影类型中出现次数最多的 5 部电影，并通过不同的颜色表示电影所属的类型，如图 6 所示。通过这样的设计，用户首先可以对用户聚类有一个总体上的概览，再通过筛选确定自己感兴趣的人或主题后，可视化系统可以将这部分人群的兴趣进行详细的展示，以供用户更进一步地进行深度分析和使用，帮助用户获取到具体的电影信息。

层次气泡图的生成过程如下：首先将用户聚类对应到中间的 4 个大圆圈，圆圈的半径与该聚类所包含的用户人数成正比；然后计算每个用户聚类对不同类型电影的偏爱程度，仅保留每个用户人群比较偏爱(偏爱值>2)的电影类型，并与用户聚类圆圈中的电影类型圆圈对应，圆圈半径与该用户聚类看过该电影类型中电影的数量成正比；再统计出每个用户聚类在每个电影类型中看得最多的 5 部电影，并将其对应于最里层的圆圈，用圆圈半径表示该用户聚类看过这部电影的次数，圆圈越大表示次数越多。最后，生成嵌套结构的数据。将数据以更加简洁轻便的 json 格式保存，以便于通过网页进行可视化展示。

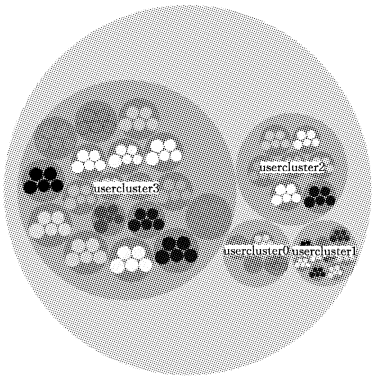


图 5　嵌套气泡图总体概览

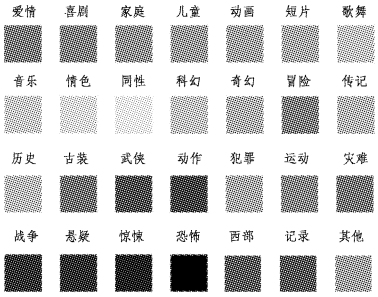


图 6　不同电影类型的颜色映射

在可视化中，我们为用户提供了方便的交互。用户如果对某一用户人群感兴趣，则可以点击代表该人群的气泡，那么在网页上就会放大该人群气泡，用户就可以观察到该人群具体感兴趣的是什么类型的电影以及它们对应的比例；在这一视图中用户还可以点击感兴趣的电影类型圆圈，那么可视化会将视角进一步放大至该圆圈，展示该类型中用户看得最多的几部电影以及比例。这种从全局到局部的布局可以让用户能从总体的概览出发进行不同人群的比较，然后根据用户的兴趣再到具体细节的交互进行可视化探索，通过用户的交互使更为详细的信息得到展示。

接下来将从用户的角度，通过可视化去交互探索地获取感兴趣的信息。

图 5 展示了嵌套气泡图的总体概览，可以看到聚类 3 的圆圈最大，说明属于聚类 3 的用户人数最多，而且他们的兴趣最为广泛，对各种类型电影的偏爱值都很高；而聚类人群 0 感兴趣的电影类型最少，只有爱情、情色、同性、奇幻，但也说明了聚类 0 的兴趣最为集中；聚类 2 的人数也比较多，它的颜色大多对应于颜色映射表中的第一行，说明聚类 0 中的用户比较喜欢温馨的电影，比如爱情、喜剧、家庭等；聚类 1 中的用户喜爱的电影颜色偏暗，说明聚类 1 中的用户口味较重，喜欢恐怖、战争、动作、犯罪等类型的电影。我们从嵌套气泡图的总体概览中观察得到的结论和热度图的基本一致，并且可以看到热度图中未能展示的一些信息，比如通过气泡的大小比较

各个用户聚类的人数。

下面将给出针对用户聚类 1 的交互展示，从而探索聚类 1 人群的兴趣细节。我们点击聚类 1 人群对应的气泡“user-cluster1”，那么视角会放大到聚类 1 人群，如图 7 所示。

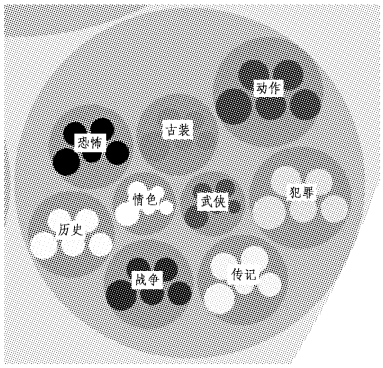


图 7 人群 1 的气泡图

通过点击聚类 1 的气泡，用户可以进一步观察聚类 1 感兴趣的具体电影类型，可以看到聚类 1 的观影口味比较重，他们喜欢动作、战争、犯罪之类的电影。若当前用户还想了解进一步的信息，比如聚类 1 比较喜欢看的犯罪片具体有哪些，那么可以在这个视图中点击“犯罪”气泡，视角会进一步放大，如图 8 所示。其展示了聚类 1 看得最多的 5 部犯罪类型的电影。之所以选择展示 5 部电影，是为了在尽可能展示更多电影的情况下保证可视化效果清晰明了，防止节点过多而带来的视觉混乱。

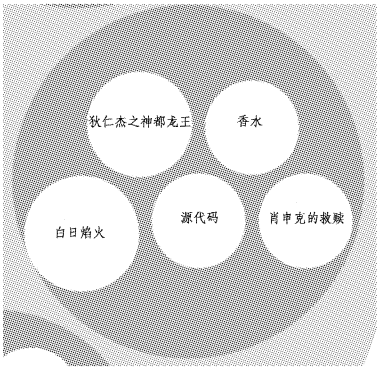


图 8 人群 1 感兴趣的犯罪片

与传统上一旦生成后就固定不变的效果图或者表格相比，层次气泡图灵活多变，可以根据用户的不同需求，通过缩放操作为用户清晰地展示数据细节，同时也可以进行整体上的宏观比较。这种层层递进的表现形式可以使用户在对所有人群有总体概览的前提下，根据自己的兴趣更深入地探索和发现更加详细的信息。

结束语 本文工作主要包括数据分析与可视化两个方面。在数据分析中，我们基于用户评分信息，通过潜在语义模型提取了用户的特征向量，并通过用户的特征向量对用户进行聚类；在可视化中，我们对聚类结果进行了图像化的映射，并通过静态的可视化和交互的可视化将聚类结果通过图形展示，从中可以发现交互地设计层次气泡图让用户可以从全局出发，根据自己的兴趣探索局部的细节，比静态的可视化更有助于聚类结果的呈现和用户的理解。

由于时间有限，文章仅仅对社交网络中的电影数据进行了可视化分析，但是在实际的社交网络中不仅仅只有电影，还有图书、音乐等。如何在一个平面空间同时对电影、图书、音乐等更多维的数据进行可视化分析，是今后工作需要解决的问题。

参 考 文 献

[1] GOU L, YOU F, GUO J, et al. Sfviz: interest-based friends exploration and recommendation in social networks[C]// Proceedings of the 2011 Visual Information Communication-International Symposium. ACM, 2011: 15.

[2] KRULWICH B. Lifestyle finder: intelligent user profiling using large-scale demographic data[J]. Artificial Intelligence Magazine, 1997, 18(2): 37-45.

[3] HOFMANN T. Latent semantic models for collaborative filtering[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 89-115.

[4] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 50-57.

[5] GOLUB G H, REINSCH C. Singular value decomposition and least squares solutions[J]. Numerische Mathematik, 1970, 14(5): 403-420.

[6] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1-3): 37-52.

[7] YE M, LIU X, LEE W C. Exploring social influence for recommendation: a generative model approach[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 671-680.

[8] CARD S K, MACKINLAY J D, SHNEIDERMAN B. Readings in Information Visualization: Using Vision To Think[M]. San Francisco: MorganKaufmann Publishers, 1999: 1-712.

[9] HERMAN I, MELANCON G, MARSHALL M S. Graph visualization and navigation in information visualization: A survey[J]. IEEE Trans. On Visualization and Computer Graphics, 2000, 6(1): 24-43.

[10] JOHNSON B, SHNEIDERMAN B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures[C]// IEEE Conference on Visualization'91. IEEE, 1991: 284-291.

[11] JOHNSON B, SHNEIDERMAN B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures[C]// IEEE Conference on Visualization'91. IEEE, 1991: 284-291.

[12] BALZER M, DEUSSEN O, LEWERENTZ C. Voronoi treemaps for the visualization of software metrics. [C]// Proceedings of the 2005 ACM Symposium on Software Visualization. New York: ACM, 2005: 165-172.

[13] FRIENDLY M. A brief history of the mosaic display[J]. Journal of Computational and Graphical Statistics, 2002, 11(1): 89-107.

[14] WANG W, WANG H, DAI G, et al. Visualization of large hierarchical data by circle packing[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2006: 517-520.

（下转第 427 页）

此次实验 Mine 类测试例与基础类的相似度均有明显回落。重复实验 2 10 次，选择 8 个类，每个类中各选取一篇文本(10 次实验选取不同的文本)与 Law 类之间计算平均相似度，并取计算结果的平均值，其结果如图 12 所示。

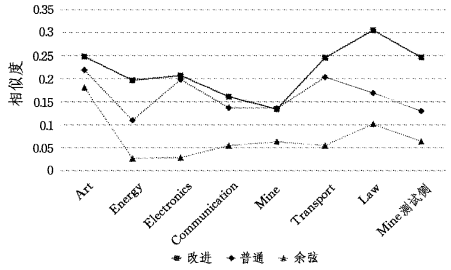


图 12　单文本与 Law 类的相似度平均值

实验结果表明，改进的语义相似度算法 TSSDWFI 很好地结合了词频与语义之间的联系，在挖掘文本之间的相似关系上具有更明显的优势，计算出的文本相似度结果更具有稳定性和准确性。

结束语　基于语义词典和词频信息的文本相似度计算与传统的方法相比，显示出了其优越性，有效地考虑了词语间的语义关系和词语间的比例关系。计算文本之间的相似度是文本挖掘过程中的一个重要步骤。现有的文本相似度计算方法大多数是利用特征空间的数值进行计算的，缺乏对词语之间语义相关性的考虑。本文在现有的词语相似度计算的基础上，提出一种改进算法，改进后的算法对文本间的相似度计算更加准确和稳定。在此基础上，可以进一步进行文本的归类以及文本中离群点的剔除工作。

参 考 文 献

[1] 陈飞宏. 基于向量空间模型的中文文本相似度算法研究[D]. 成都：电子科技大学，2011.

[2] 张振亚，王进，程红梅，等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学，2005，32(9)：160-163.

[3] 吴奎，周献中，王建宇，等. 基于贝叶斯估计的概念语义相似度算法[J]. 中文信息学报，2010，24(2)：52-57.

[4] 郭庆琳，李艳梅，唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究，2008，25(11)：3256-3258.

[5] 卫驰. 基于 TFIDF 的文本分类算法[D]. 杭州：浙江大学，2015.

[6] 冯荣俊. 基于文档频率的特征提取算法的改进及应用[D]. 南京：南京邮电大学，2005.

[7] 韩如冰，叶得学. 基于 VSM 的权重改进文档相似度算法研究[J]. 软件，2012，33(10)：103-105.

[8] 王格，吴钊，李向. 基于全文检索的文本相似度算法应用研究[J]. 计算机与数字工程，2016，44(4)：567-571.

[9] 刘杰，郭宇，汤世平，等. 基于《知网》2008 的词语相似度计算[J]. 小型微型计算机系统，2015，36(8)：1728-1733.

[10] 吴健，吴朝晖，李莹，等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报，2005，28(4)：595-602.

[11] 张沪寅，刘道波，温春艳. 基于《知网》的词语语义相似度改进算法研究[J]. 计算机工程，2015，41(2)：151-156.

[12] 肖志军，冯广丽. 基于《知网》义原空间的文本相似度计算[J]. 科学技术与工程，2013，13(29)：8651-8656.

[13] 孙润志. 基于语义理解的文本相似度计算研究与实现[D]. 辽宁：中国科学院研究生院(沈阳计算技术研究所)，2015.

[14] 袁晓峰. 基于《知网》的文本相似度研究[J]. 成都大学学报(自然科学版)，2014，33(3)：251-253.

[15] 陈攀，杨浩，吕品，等. 基于 LDA 模型的文本相似度研究[J]. 计算机技术与发展，2016，26(4)：82-85.

[16] 王蒙. 基于 LDA 的文本推荐算法的研究及在文献检索的应用[D]. 沈阳：辽宁大学，2015.

[17] 王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学，2013，40(12)：229-232.

[18] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版)，2010，28(6)：603-608.

[19] 杜坤，刘怀亮，王帮金. 基于语义相关度的中文文本聚类方法研究[J]. 情报理论与实践，2016，39(2)：129-133.

[20] 《同义词词林扩展版》[EB/OL]. http://www.ir-lab.org.

[21] AGIRRE E，RIGAU G. A Proposal for Word Sense Disambiguantion Using Conceptual Distance[C]// Proc. of Recent Advances in NLP(RANLP). 1995：258-264.

[22] 李荣陆. Reuters-21578 语料说明[EB/OL]. http://more.datatang.com/data/43318.

(上接第 390 页)

[15] KEIM D，ANDRIENKO G，FEKETE J D，et al. Visual analytics：Definition，process，and challenges[M]. Springer Berlin Heidelberg，2008.

[16] PIKE W A，STASKO J，CHANG R，et al. The science of interaction[J]. Information Visualization，2009，8(4)：263-274.

[17] KAUFMAN L，ROUSSEEUW P J. Finding groups in data：an introduction to cluster analysis[M]. John Wiley & Sons，2009.

[18] EICK S G. Graphically displaying text[J]. Journal of Computational and Graphical Statistics，1994，3：127-142.

[19] STASKO J. Information visualization[OL]. http://www.cc.gatech.edu/ classes /AY2004/cs7450_spring.

[20] FENG Y D，WANG G P，DONG S H. Information Visualization[J]. Journal of Engineering Graphics，2001：324-329.

[21] SMITH M A，SHNEIDERMAN B，MILIC-FRAYLIN N，et al. Analyzing (social media) networks with NodeXL[C]// Proceedings of the Fourth International Conference on Communities and Technologies. ACM，2009：255-264.

[22] HENRY N，FEKETE J D. MatrixExplorer：a Dual-Representation System to Explore Social Networks[J]. IEEE Transactions on Visualization & Computer Graphics，2006，12(5)：677-684.

[23] FRUCHTERMANN T M J，REINGOLD E M. Graph drawing by force-directed placement[J]. Software：Practice and experience，1991，21(11)：1129-1164.