基于信息内容和拓扑关系的社会媒体用户兴趣分类

吴海涛[1,2] 应 时[1]
(武汉大学计算机学院软件工程国家重点实验室 武汉 430072)[1]
(黄淮学院软件学院 驻马店 463000)[2]

摘 要 随着社会的发展,信息已经成为社会发展越来越重要的部分,人类的信息传播活动越来越明显地展示出分众特征,对用户的分类成为人类信息活动的一个重要研究课题。从这一目标出发,分别基于信息内容、拓扑关系和两者综合的方法,按兴趣主题对社会媒体用户进行分类。对于基于信息内容的用户分类,采用LDA主题模型从用户所发布的内容中提取其主题分布,基于这一分布,采用支持向量机、决策树、贝叶斯等多种模型按兴趣主题对用户进行分类。对于基于拓扑关系的分类,依据相同兴趣主题的用户倾向于拥有共同的粉丝这一发现,构建分类模型来按兴趣主题对用户进行分类。然后提出综合信息内容和拓扑关系的分类方法来对用户进行分类。最后基于大规模Twitter数据的实验发现,采用综合方法对用户进行的兴趣分类性能明显高于采用单一信息内容或粉丝拓扑方法的性能。

关键词 在线社会网络,兴趣分类,LDA,粉丝拓扑
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.037

Classifying Interests of Social Media Users Based on Information Content and Social Graph

WU Hai-tao[1,2] YING Shi[1]
(State Key Laboratory of Software Engineering, Computer School, Wuhan University, Wuhan 430072, China)[1]
(Software College, Huanghuai University, Zhumadian 463000, China)[2]

Abstract With the development of society, there has been a more and more obvious presence of the characteristic of audience-segmentation in human activity over information spreading, and user classification has also become an important research topic. So the article carried out a study over online social network user from multiple perspectives which mainly include user classification based on interested topics and preference, classify interests of social media user based on information content and topological relation, and both them respectively. For user classification based on information content, we adopted LDA to extract the topic distribution from the content posted by users. And the distribution is used in support vector machine, decision tree, Bayes and other multiple models to classify interests of users. For user classification based on topological relation, we found that users with same interests tend to have more common fans, and based on this finding we built classification models to classify users. Then, we proposed methods of combining information content and topological relation to classify users. Based on the experiments using Twitter data, we found that the combined method outperforms the one based on information content or topological relation.

Keywords Online social networks, User classification, LDA, Topological relation

1 引言

Twitter是目前最为流行的在线社会网络平台之一,也是世界上最具影响力的微博社区,全球有十多亿人通过这一平台分享自己生活或工作中的各种经验和感受,同时通过这一平台维持朋友或其他社会关系,并不断扩展自己的朋友圈和社交网络。当用户通过Twitter表达自己的兴趣爱好或对某一主题的关注时,也为网站经营者提供了机遇,如发布针对性的广告内容获得经济收益,或者更好地提供个性化服务吸引更多的用户等。如果能够对用户进行准确的分类,则可以在很大程度上提高广告推广和朋友推荐的效果,增加信息传播的有效性,这具有重要的应用价值,因此进行了大量针对在线社会网络用户分类的研究成果。

本文首先介绍兴趣分类的相关研究工作;然后提出基于用户微博内容的兴趣分类、基于用户粉丝拓扑的兴趣分类和综合用户微博内容和粉丝关系的兴趣分类方法;最后给出实验验证。

2 相关研究

对在线社会网络用户的分类研究大多是以Twitter为研

到稿日期:2014-04-27 返修日期:2014-08-11 本文受国家自然科学基金项目(61070012,61070022),国家自然科学基金重点项目(91118003,61272113,61272108)资助。

吴海涛(1974—),男,博士,副教授,CCF会员,主要研究领域为面向服务软件工程、智能信息处理;应 时(1965—),男,教授,博士生导师,CCF高级会员,主要研究领域为面向对象软件工程方法、软件体系结构和模式、软件的可重用性与互操作性。

究对象，依据不同标准对用户进行分类。这不仅因为 Twitter 自身拥有巨大的影响力，还因为 Twitter 同时具有社交网络和社会媒体的功能，大量的各类媒体、公司和个人等都通过这一平台发布或获取信息，因此如何识别 Twitter 帐户是机构或个人，还是具有哪些特点的用户类别，也就吸引了众多研究者的目光。Choudhury 等[1]采用标准机器学习框架将 Twitter 用户分为组织、记者媒体和普通用户 3 个类别。Perez-Sola等[2]对在线社会网络用户进行两种方式的分类，一种是根据拓扑结构相关的特征建立分类器将用户分为公司和个人，另一种是根据结点度、分族系数和用户关系建立分类器将用户分为博主、明星、媒体、组织和普通用户。Chu 等[3]研究如何判断 Twitter 帐户是由人(humans)、僵尸(bots)、人机协助(cyborgs)来进行发布消息、关注其他用户等操作。他们首先分析这 3 类帐户的区别，然后基于分析结果提取行为和内容等相关特征构建分类器，对 Twitter 帐户进行分类。

除了识别用户是机构还是个人外，研究者们还对 Twitter 用户依据兴趣或政治倾向等进行分类。Pennacchiotti 等[4]提出一个综合的架构对社会媒体用户进行分类，这一架构同时考虑用户相关的特征以及拓扑相关的特征等对用户进行分类，基于 Twitter 数据，分别对用户的政治倾向、种族类别和是否为某一用户的粉丝进行了分类，并取得了较高的准确率。何炎祥等[5]对微博用户进行了分类，他们提出时间片微元的概念，并建立了时间片微元模型，然后对每个时间片内的微博所涉及到的用户进行研究，得到时间片微元内部的用户兴趣度向量，最终整合所有时间片内的用户兴趣度向量，再对整个时间段内用户的兴趣度向量进行两次朴素贝叶斯分类，得到整个时间段内的用户分类；基于新浪微博数据的实验表明，所提出的方法能有效对大规模的微博语料中所涉及到的用户进行较准确的分类。

本文首先基于用户所发布的微博内容对用户按兴趣进行分类，即采用 LDA 主题模型从微博内容提取用户的兴趣主题分布，基于这一主题分布采用支持向量机、决策树、贝叶斯等多种模型对用户按兴趣进行分类；然后基于粉丝拓扑关系对用户按兴趣进行分类，基于具有相同兴趣主题的用户倾向于拥有共同的粉丝的发现，构建分类模型对用户进行分类；最后综合上述两种方法进行分类，进一步提高准确率。

3 数据描述

为了获取更为理想的 Twitter 用户样本，这里从一个非常流行的提供 Twitter 用户列表的网站 Twellow(www.twellow.com)上收集 Twitter 用户列表。Twellow 网站不仅自动地从 Twitter 网站上收集微博内容的发布者帐户信息，还吸引了很多 Twitter 用户，把自己的 Twitter 帐户信息手工添加到这一网站，因此能够得到相对全面的用户信息。

从 Twellow 网站上获取的用户列表中的用户并不完全符合实验需求，需要进行数据过滤，即过滤掉缺少实验必需信息的用户，比如可能因为帐户停用或设置隐私保护而无法通过 Twitter APIs 获取个人简介信息的用户、最近 3 个月内没有发布过任何微博的用户、语言设置为非英语的用户，前两种用户被过滤掉的原因是缺少分类需要的依据信息，而非英语用户被过滤掉则是因为本文的研究需要采用主题模型对微博内容进行分析，这需要用户的微博文字必须为同一种语言，而在 Twitter 中，英语是最为流行的语言[6]。经过数据收集和过滤，最终获得了可用于实验的一百多万个 Twitter 帐户及其相关信息。

对于所获取的这些有效用户，文中使用 Twitter REST APIs 获取每个用户的个人简介、全部粉丝列表，以及他们的 Lists。同时为了使用微博内容进行兴趣分类，同样使用 APIs 获取这些用户所发布的微博内容，对于每一个用户获取他们在 2013 年 9 月 1 号之前的 200 条最新发布的微博，对于总微博数量不足 200 的用户，则获取他们所有的微博内容。表 1 显示了本文所用数据的详细信息。

表 1 微博数据描述

Number of users	1.562M
Number of lists links	28.76M
Number of follower links	5837.48M
Number of tweets	301.25M

4 用户兴趣类别标准

对本文分类方法进行实验验证，使用基于 Twitter Lists 对用户分类的结果作为用户分类的参照标准。

Twitter Lists 是 Twitter 公司提供的用来让用户管理自己关注的用户列表的一项新功能，主要目的是让用户可以将自己所关注的同类别的用户放在同一个 List 中，以方便查找所关注用户的信息或浏览所关注用户的微博。用户可以自由地创建 Lists 并对它们进行命名，将自己所关注的用户添加在适当的 Lists 进行管理。

采用 Twitter Lists 对用户分类主要是基于充分使用群体智慧(Wisdom of the crowd)的思想，即如果某一用户被高频率地放入具有类似名称的 Lists 中时，这一用户将可以认为是大众公认的属于与这些 Lists 名称相近的兴趣类别，例如 CNN 的帐户被大量的用户放入具有“News”名称的 Lists 中，那么 CNN 就被认为属于 News 兴趣类别。根据这一思想，首先对每个用户所在的 Lists 的名称进行合并统计，如把 Media 归到 News 类等，并计算出用户 Lists 名称的频率，然后再根据出现频率最高的 Lists 名称来对用户进行归类。最终根据 Twellow 网站上的类别流行情况将用户分为 9 个兴趣类别：新闻(News)、音乐(Music)、科技(Tech)、体育(Sports)、饮食(Food)、政治(Politics)、健康(Health)、教育(Education)和旅游(Travel)。最后计算每个用户涉及的 Lists 频率最高的名称，与这 9 个兴趣类别进行匹配，根据匹配结果把用户放入相应的兴趣类别。

5 用户兴趣分类

5.1 基于用户微博内容的兴趣分类

用户所发布的微博内容直接地反映他们的兴趣偏好，因为用户发布微博的数量庞大，如果直接使用全部微博内容对用户兴趣进行分类，大量的信息内容会严重地影响计算效率，直接导致计算成本增加。为了能够更快速准确地使用微博内容对用户依据兴趣偏好进行分类，本文采用 LDA 主题模型来提取用户微博内容的主题分布，然后使用提取出的主题分布对用户进行分类。

5.1.1 主题分布提取

根据已有的研究，转载和包含短网址的微博能更准确地反映用户的兴趣偏好[7]，而回复和包括 Hashtag 的微博与用户的兴趣主题相对关系较弱，所以只选用转载和包含短网址的微博作为主要数据源来提取用户兴趣主题。本文将每个用户所发布的微博内容合并成一个文档，将其作为 LDA 模型的输入，模型根据这一输入以及设定的相关参数，产生每个用户的主题分布。

类似文献[7]，在使用主题模型处理微博内容之前，先对微博内容进行相关的数据清理，比如移除 420 个 stop-words 和不在 Wikipedia 数据集中的术语，以及去除出现次数小于 10 的单词；然后将这些处理过的微博内容输入主题模型，参考文献[8]，假设主题数量 N 为 50，超参数 α 设为 1，β 设为 0.01。采用 Gibbs 抽样算法[9]近似估算进行 2000 次迭代后每个用户的主题分布。那么每一个用户的主题分布共 50 个值，每个值代表着这一主题在这 N 个主题中所占的百分比，当然每个用户的 50 个主题分布值之和等于 1。

5.1.2 基于微博内容进行用户兴趣分类

为了分析预测模型对预测性能的影响，同时获取更高的准确率，本文采用支持向量机（Support Vector Machine，SVM）、Random Forests、Bagging、J48 decision tree 和 Naive Bayes 等 5 种分类模型。对于支持向量机模型，采用比较常用的 LIBSVM[10]，实现了支持向量机模式识别与回归的功能，其中采用 Radial Basis Function（RBF）核函数；其他模型采用 Hall 等[11]开发的机器学习集成模块 WEKA 进行实验。这几种模型均采用十折交叉验证法（10-fold cross-validation）。通过比较准确率（Accuracy）、正确率（Precision）和 F 值（F-measure）来评价预测性能。准确率是判定正确的结果数量占总体数量的比例，反映了模型对整体样本的判定能力。

预测结果如表 2 所列，从表中可知各种模型的预测结果总体较好，在准确率上，最好的结果是支持向量机模型，最差的是朴素贝叶斯模型，其差值超过 14%。这说明模型对预测结果有一定的影响，在采用 LDA 对用户兴趣分类方面要充分考虑预测模型的优劣。无论在准确率还是 F 值上，支持向量机模型都取得了最好的预测结果，其正确率达 78.26%，F 值达 77.94%，这说明微博内容可以较准确地反映他们的兴趣，采用支持向量机模型对用户兴趣进行分类可以达到 78% 的准确率。

表 2 基于微博内容的用户兴趣分类

Models	Accuracy (%)	Precision (%)	F-measure (%)
SVM	78.26	78.13	77.94
Bagging	75.43	74.93	75.10
Random Forests	74.18	73.78	73.67
J48 Decision Trees	68.25	68.25	68.24
Naive Bayes	64.51	69.81	63.82

5.2 基于用户粉丝拓扑的兴趣分类

要采用拓扑链接按兴趣对用户进行分类，需要知道 Twitter 用户粉丝关系组成的网络与其兴趣类别之间的关系。当前研究表明，在线社会网络广泛地存在着同质性现象，例如 Hofman 等[12]发现在 Twitter 中的粉丝关系上存在着同质性，即用户更倾向于关注同类别的用户。依据这些研究所得出的结果，假设 Twitter 中粉丝关系与其兴趣偏好之间存在同质性，即具有相同兴趣的用户倾向于拥有共同的粉丝。如果这一假设能够被证实，那么就可以根据一个用户的粉丝与已知兴趣类别的用户粉丝的交集情况来推断出此用户所属的兴趣类别，从而实现准确的分类。

5.2.1 相同兴趣类别用户粉丝一致性

本节将主要对下述假设进行验证：具有相同兴趣的用户更倾向于拥有共同的粉丝。这一假设的验证通过比较两个用户共同粉丝数量进行，也就是说只要证明同一类别的两个用户的共同粉丝数量普遍高于不同类别的两个用户的共同粉丝数量，即可说明这一假设是成立的。

为了验证这一假设，我们对不同兴趣类别的用户粉丝交集情况进行了比较分析。首先根据用户的兴趣类别把用户放入不同的组，然后计算任意两个类别间的用户粉丝交集元素个数的平均值，具体统计结果如表 3 所列。

表 3 类别间用户粉丝交集元素平均个数

Category	Food	Health	Music	News	Education	politics	Sports	Tech	Travel
Food	21.47	7.18	3.13	7.32	1.73	6.20	3.15	5.09	8.15
Health	7.18	36.89	4.38	12.57	3.65	9.63	3.16	8.35	8.24
Music	3.13	4.38	22.77	6.92	1.24	10.79	8.72	4.41	3.97
News	7.32	12.57	6.92	20.48	3.72	17.54	6.46	10.58	9.86
Education	1.73	3.65	1.24	3.72	12.39	3.30	1.40	2.81	2.39
politics	6.20	9.63	10.79	17.54	3.30	62.29	9.92	10.00	7.14
Sports	3.15	3.16	8.72	6.46	1.40	9.92	45.05	4.48	3.68
Tech	5.09	8.35	4.41	10.58	2.81	10.00	4.48	15.04	6.69
Travel	8.15	8.24	3.97	9.86	2.39	7.14	3.68	6.69	62.35

表 3 显示了各类别间用户粉丝交集元素个数的平均值，表中的任一数字表示对应的两个类别的用户粉丝交集元素个数的平均值。例如表中第 2 行第 2 列的数字 21.47 表示 Food 类别与 Food 类别中用户粉丝交集元素个数的平均值，第 2 行第 3 列的数字 7.18 表示 Food 类别中用户与 Health 类别中用户粉丝交集元素个数的平均值。由表可知，同类别的用户粉丝交集元素个数总是远远高于与其他类别的用户粉丝交集元素个数。其中同类别用户粉丝交集与不同类别粉丝交集元素个数差别最大的为 Travel 类别，Travel 同类别的值 62.35 比 Travel 类别与（次高点的）News 类别的值 9.86 高出 6 倍。

为了显示更为直观的结果，图 1 示出表 3 中数字标准化后的结果，图中用灰度表示两个类别中用户粉丝交集元素个数平均值的大小，颜色越深对应两个类别间的粉丝交集越大，反之则越小。由图可知，颜色最深的方块始终处于图的对角线上，这意味着相同兴趣类别的用户粉丝交集元素个数总是最大的。由表 3 和图 1 可知，具有相同兴趣类别的用户粉丝之间交集元素个数远远多于具有不同兴趣类别的用户粉丝交

集元素个数，即具有相同兴趣类别的用户倾向于拥有共同的粉丝，这一假设是成立的。

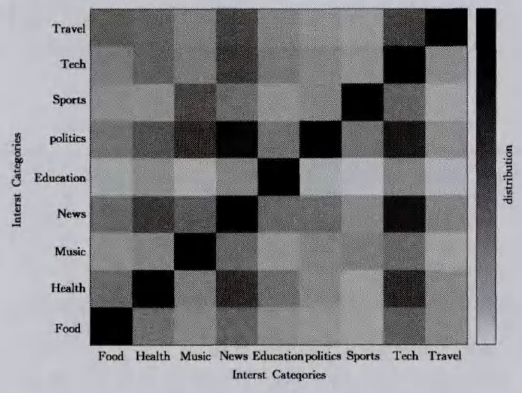

图 1　类别间用户粉丝交集元素平均个数分布

以上分析说明粉丝关系可以反映出其兴趣类别情况，即一个用户与同类别的用户粉丝交集元素个数较多，而与不同类别的用户粉丝交集元素个数较少。例如对于一个用户 u_i，如果其属于 Food 类别，那么根据表 3 的结果，其与 Food 类别的用户粉丝交集元素个数会明显多于与其他类别用户的粉丝交集元素个数。根据这一特性，则可以采用粉丝关系识别出用户所属的兴趣类别，即根据待分类用户与已知类别用户的粉丝交集大小，可以将待分类用户标记为粉丝交集数量较大的用户所在的兴趣类别。比如对于一个待分类用户 u_j，如果其与 Food 类别的用户粉丝交集元素平均值大于与其他类别的平均值，那么这一用户很可能属于 Food 类别。

5.2.2　基于粉丝关系进行用户兴趣分类

从以上分析可以发现，用户的粉丝数量与他们所在的兴趣类别存在明显的关系。下面具体介绍如何根据用户的粉丝关系对用户进行兴趣分类。因为具有共同兴趣偏好的用户粉丝交集较大，所以可以将未知兴趣类别的用户粉丝与已知兴趣类别的用户粉丝进行比较，根据比较结果来推断未知类别的用户类别归属。

为了提高分类准确率，依据上述关系，本文把数据集中的用户随机地平均分成 3 部分：基础集（base set）、训练集（training set）和测试集（test set）。其中基础集为已知兴趣类别的用户集，主要用于计算其用户（训练集和测试集中的用户）的粉丝与各兴趣类别中用户的粉丝的交集个数；训练集也为已知兴趣类别的用户集，用于构建分类模型（如 SVM、J48 决策树等）；测试集则用于检验分类模型的性能。

在采用分类模型对用户进行兴趣分类之前，首先要根据基础集中的用户兴趣主题信息来计算训练集和测试集中所有用户的粉丝与每个兴趣类别的用户的粉丝交集个数。下面以测试集为例来说明计算交集个数的方法。

假设测试集中用户个数为 m，则第 $i(1\leqslant i\leqslant m)$ 个用户为 u_i^{test}，其粉丝集合为 s_i^{test}；基础集用户个数为 n，第 $k(1\leqslant k\leqslant 9)$ 个兴趣组中用户个数为 n_k，第 k 个兴趣组中第 $j(1\leqslant j\leqslant n_k)$ 个用户为 u_{kj}^{base}，其粉丝集合为 s_{kj}^{base}。则为了对用户 u_i^{test} 进行兴趣分类，需要计算 u_i^{test} 的粉丝和兴趣组 k 中每个用户的粉丝交集元素个数的平均值 A_i^k：

$$A_i^k=\frac{1}{n_k}\sum_{j=1}^{n_k}|s_i^{test}\bigcap s_{kj}^{base}| \tag{1}$$

式中，$|s_i^{test}\bigcap s_{kj}^{base}|$ 指两个用户的粉丝集合的交集元素个数。由于本文把用户分为 9 个兴趣类别，对于任一用户 u_i^{test}，可以根据式(1)计算出 9 个兴趣类别对应的平均值 $(A_i^1,A_i^2,\cdots,A_i^k,\cdots,A_i^9)$，每个值代表着这个用户与相应的兴趣类别的接近程度。直观地，可以把此用户标记为平均值 A_i^k 最大的值所对应的兴趣类别。但为了增加分类的准确率，这里对训练集和测试集中的每个用户都计算出 9 个兴趣类别的平均值，然后把这 9 个值作为特征值，用训练集中的数据构建分类模型，再用测试集中的数据对分类模型进行性能评估。

式(1)直接使用两个用户粉丝的交集个数来判断 u_i^{test} 的兴趣类别，称为直接交集法（direct interaction）。然而这种方法在用户的粉丝数量差别较大的情况下可能会造成一定误判，比如对于测试集中的一个用户 u_i^{test}，设其真实的兴趣类别是 travel，然而当基础集中的某一兴趣类别（如 news）中大部分用户的粉丝数量远大于 travel 类别时，采用直接交集法很可能将用户 u_i^{test} 判定为 news，这是因为 news 中用户粉丝数量非常大，用户 u_i^{test} 的粉丝则与其交集的元素个数可能会因为基数大而较多，从而导致这一错误判定结果。为了避免这种情况，本文分别分析了使用交集个数除以最大、最小和平均粉丝数量的情况，相应地，这 3 种方法称为最大（max interaction）、最小（min interaction）、平均（avg nteraction）交集法，其对应的计算公式分别如下：

$$A_i^k=\frac{1}{n_k}\sum_{j=1}^{n_k}\frac{|s_i^{test}\bigcap s_{kj}^{train}|}{\max(|s_i^{test}|,|s_{kj}^{train}|)} \tag{2}$$

$$A_i^k=\frac{1}{n_k}\sum_{j=1}^{n_k}\frac{|s_i^{test}\bigcap s_{kj}^{train}|}{\min(|s_i^{test}|+1,|s_{kj}^{train}|+1)} \tag{3}$$

$$A_i^k=\frac{1}{n_k}\sum_{j=1}^{n_k}\frac{|s_i^{test}\bigcap s_{kj}^{train}|}{(|s_i^{test}|+|s_{kj}^{train}|)/2} \tag{4}$$

式(3)中分母为两个粉丝集合个数的最小时，会出现分母为零的情况，为了避免这种情况，将每个粉丝集合的数量加 1。

表 4　基于粉丝交集的用户兴趣分类

Method	Models	Accuracy (%)	Precision (%)	F-measure (%)
direct interaction	SVM	76.68	75.47	76.52
	Bagging	78.97	79.24	79.06
	Random Forests	**79.24**	**79.51**	**79.32**
	J48 Decision Trees	78.06	78.01	78.04
max interaction	SVM	75.76	76.02	75.29
	Bagging	**77.29**	**76.93**	**76.58**
	Random Forests	76.52	76.35	75.24
	J48 Decision Trees	76.01	76.84	76.63
min interaction	SVM	77.23	76.58	77.61
	Bagging	**81.42**	**82.31**	**82.16**
	Random Forests	79.05	79.25	79.43
	J48 Decision Trees	78.14	78.21	78.07
avg interaction	SVM	76.79	76.04	76.46
	Bagging	**78.84**	**79.11**	**79.23**
	Random Forests	77.26	77.51	77.32
	J48 Decision Trees	78.16	77.91	78.29

针对以上 4 种求交集的方法，首先使用训练集来构建

SVM、Random Forests、Bagging、J48 decision tree 等 4 种分类模型(由于采用 Naive Bayes 模型的分类准确率较低,因此这里不再考虑该模型),然后用测试集来评估其性能。实验工具及具体方法同第 5.1.2 节,分类结果如表 4 所列,每种方法的最好的结果(最高的准确率)用黑体字符显示。

从表中可知,采用不同的求交集方法产生了不同的分类性能,其中最大交集法获得的结果最差,当采用 Bagging 模型时准确率为 77.29%,最小交集法获得的结果最好,采用 Bagging 模型的分类准确率达到 81.42%。采用粉丝关系的最好分类结果比采用微博内容的准确率 78.26%(见表 2)稍高一些,这说明根据用户的粉丝关系可以很有效地确定用户兴趣类别。另外对于同一种求粉丝交集的方法,采用的分类模型对分类性能的影响较小,不同分类模型所取得的分类准确率差别不大。

5.3 综合用户微博内容和粉丝关系的兴趣分类

本节采用比较概率估计值的方法对前面所述的两种分类方法进行综合,即对于每一个用户,采用其获得的概率估计值最高的分类方法所得到的结果。概率估计是分类模型产生的,表示一个用户属于某一兴趣类别的可能性大小,其值在 0 和 1 之间。在本节的实验中,任一用户在一个分类模型中都有 9 个概率估计值,即表示这一用户属于 9 个兴趣类别的概率,分类模型会把这一用户标记为具有最高概率估计的兴趣类别。分别用 $prob^{lda}(u_i)$ 和 $prob^{follower}(u_i)$ 表示采用基于 LDA 的微博内容和用户粉丝关系 u_i 的分类模型中用户获得的最高概率估计值。以此为基础,可以根据下面的公式来确定用户 u_i 使用哪种分类模型的分类结果。

$$score(u_i)=\alpha \cdot prob^{lda}(u_i)-(1-\alpha) \cdot prob^{follower}(u_i) \quad (5)$$

式中,α 为调节因子,范围为[0,1],用于说明各分类模型所占的权重,当 α 为 1 时表示只使用基于 LDA 的微博内容对用户进行兴趣分类,当 α 为 0 时表示只使用粉丝关系。实际应用中 α 可根据两种分类模型的准确率进行设定,即当使用基于 LDA 的微博内容的分类模型准确率较高时,α 的值应大于 0.5,表示此模型所占比重大;反之 α 的值应小于 0.5,表示此模型所占比重小。当 $score(u_i)$ 的值大于 0 时,表示用户的分类结果采用基于 LDA 的微博内容模型的;当 $score(u_i)$ 的值小于 0 时,表示用户的分类结果采用粉丝关系模型。

为了检验综合方法的性能,首先把数据集中的用户随机地平均分成 3 部分:基础集、训练集和测试集,并按第 5.2.2 节所述方法计算出训练集和测试集中每个用户与基础集中每个类别用户的粉丝交集元素个数。这样基于微博内容和粉丝关系的分类方法就可以使用训练集来构建分类模型。对于基于微博内容的分类,采用表 2 中性能最好的支持向量机的模型。对于基于粉丝关系的分类,同样也采用表 4 中性能最好的基于 min interaction 方法的 Bagging 模型。对于测试集中的任一用户 u_i,提取他在两种预测模型中概率估计的最大值 $prob^{lda}(u_i)$ 和 $prob^{follower}(u_i)$,采用式(5)来决定这一用户最终采用哪个模型的预测结果。

因为公式中 α 值会影响最终分类结果,所以分别测试了 α 取不同值时分类结果的准确率,其结果如图 2 所示。从图中可知,当 α 为 0.45 时,准确率达到最高 87.42%,以此值为中心,随着 α 的增大和减小,准确率明显降低。这是由于采用微博内容和粉丝关系的两种分类方法的准确率相差不大,前者为 78.26%,后者为 81.42%。所以当 α 在 0.5 附近时两种方法的综合准确率达到峰值,同时因为采用粉丝关系的分类方法准确率稍高于采用微博内容的分类方法,所以 α 小于 0.5 即向采用粉丝关系的方法倾斜时才达到最好的效果。另外图中的最高准确率 87.42%比采用基于微博内容和粉丝关系的最高准确率 81.26%高出约 6%,这说明综合分类方法进一步提高了分类准确率。

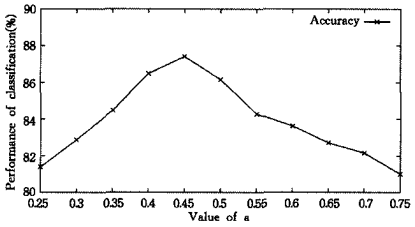

图 2 比较概率估计值的用户兴趣分类

结束语 本文提出多种方法对在线社会网络用户按兴趣进行了分类;首先依据用户所发布的微博内容采用 LDA 主题模型对用户分类;然后基于粉丝拓扑关系对相同兴趣的用户倾向于拥有共同的粉丝对用户分类;最后综合上述两种方法,提出一种比较概率估计值的方法来综合基于信息内容特征和粉丝关系拓扑的方法,即对于每一个用户,采用其获得的概率估计值最高的分类方法所产生的结果,综合方法明显地提高了分类效果。

参 考 文 献

[1] Choudhury M D, Diakopoulos N, Naaman M. Unfolding the event landscape on twitter: classification and exploration of user categories[C]// Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. 2012: 241-244

[2] Perez-Sola C, Herrera-Joancomarti J. Classifying online social network users through the social graph[C]// Proceedings of the 5th international conference on Foundations and Practice of Security. 2012, 115-131

[3] Chu Z, Gianvecchio S, Wang H, et al. Who is tweeting on Twitter: human, bot, or cyborg? [C]// Proceedings of the 26th Annual Computer Security Applications Conference. 2010: 21-30

[4] Pennacchiotti M, Popescu A-M. Democrats, republicans and starbucks afficionados: user classification in twitter[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011: 430-438

[5] 葛红美,何炎祥,陈强,等. 一种基于时间片的微博用户分类方法[J]. 小型微型计算机系统,2013(11):2441-2445

[6] An Exhaustive Study of Twitter Users Across the World-Beevolve, Social Media Analytics Platform[EB/OL]. http://www.beevolve.com/twitter-statistics/

[7] Xu Z, Ru L, Xiang L, et al. Discovering User Interest on Twitter with a Modified Author-Topic Model[C]// Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Volume 01, 2011: 422-429

[8] Zhang C, Sun J. Large scale microblog mining using distributed MB-LDA[C]// Proceedings of the 21st International Conference Companion on World Wide Web LSNA Workshop. 2012: 1035-1042

(下转第 198 页)

为 EMASK、BEMASK 和 Apriori 3 种算法在处理同一种数据库不同支持度情况下的时间比较图。图 5 为 BEMASK 相对于 EMASK 的效率随着支持度的变化趋势图。

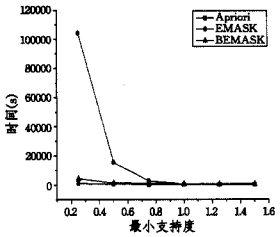

图 4 基于 T25I4D1MN1K 数据库的 3 种算法的时间消耗

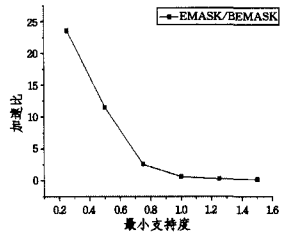

图 5 基于 T25I4D1MN1K 数据库的 EMASK 与 BEMASK 执行时间比

表 5 为基于 T50I8D1MN1K 数据库的 EMASK、BEMASK 两种算法在不同支持度(0.25%、0.5%、0.75%、1.0%、1.25%、1.5%)情况下的处理时间对比。图 6 为 EMASK 和 BEMASK 两种算法在处理同一种数据库不同支持度情况下的时间比较图。图 7 为 BEMASK 相对于 EMASK 的效率随着支持度的变化趋势图。

表 5 BEMASK 与 EMASK 算法执行时间记录

数据集	扭曲参数		支持度	EMASK	BEMASK
	p	q	(%)		
T50I8D1MN1K	0.5	0.97	1.0	67901	2075
			1.5	23417	995
			2.0	10069	621
			2.5	4048	549
			3.0	1196	435

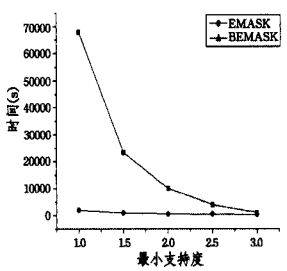

图 6 基于 T50I8D500KN1K 数据库的 EMASK 与 BEMASK 的时间消耗

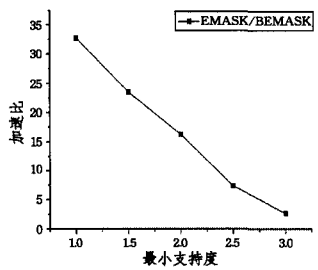

图 7 基于 T50I8D500KN1K 数据库的 EMASK 与 BEMASK 执行时间比

通过实验的结果我们可以得出如下结论：

(1)BEMASK 算法的效率相对于 EMASK 来说有很大幅度的提高。最小支持度越小，BEMASK 消耗的时间与 EMASK 消耗的时间差距越大。

(2)当数据库变得稠密后，BEMASK 的优势更加明显。

(3)由于在 BEMASK 算法中增加了一个将数据库文件转化成 Bitmap 文件的过程，因此最小支持度大于一定点时，3-项以上的频繁集大幅度减少，这样使用 Bitmap 技术来提高速度的优势就得不到明显的体现。

(4)由于增加了隐私保护过程，因此 BEMASK 算法比没有进行隐私保护过程的 Apriori 算法效率稍低。

结束语 隐私保护数据挖掘一直是数据挖掘领域一个重要的研究方向。本文针对 EMASK 算法所依据的 Apriori 算法需要完全的数据库扫描并且进行多次比较操作来降低效率的弊端，提出了 BEMASK 算法，该算法利用粒度方式将关系数据表转换成面向机器的关系模型，将数据处理转换成粒度计算的方式，计算频繁项集变成了计算基本颗粒的交集。实验证明，在保证准确性不降低的情况下，相对 EMASK 算法，算法中所采用的数据垂直 Bitmap 表示(vertical Bitmap)减少了 I/O 操作的次数，效率得到了较大的提高。

参 考 文 献

[1] 王艳，乐嘉锦，孙捷，等. 网络用户行为的隐私保护数据挖掘方法[J]. 计算机工程与应用，2012，48(13)：138-143

[2] 马进，李锋，李建华. 分布式数据挖掘中基于扰乱的隐私保护方法[J]. 浙江大学学报：工学版，2010，44(2)：276-282

[3] 张鹏，童云海，唐世渭，等. 一种有效的隐私保护关联规则挖掘方法[J]. 软件学报，2006，17(8)：1764-1774

[4] 孙茂华. 安全多方计算及其应用研究[D]. 北京：北京邮电大学，2013

[5] Pawlak Z, Grzymala-Busses J, Slowinski R, et al. Rough sets [J]. Communications of the ACM, 1995, 38(11): 88-95

[6] Zadeh L A. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems [J]. Soft Computing, 1998, 2(1): 23-25

[7] Yao Yi-yu. The Art of Granular Computing[C]// Proc of the International Conference on Rough Sets and Emerging Intelligent Systems Paradigms, 2007. Warsaw, Poland, 2007: 101-112

[8] Chen Hong-mei, Li Tian-rui, Ruan Da, et al. A rough-set based incremental approach for updating approximations under dynamic maintenance environments [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(2): 274-284

[9] 王磊，李天瑞. 一种基于矩阵的知识粒度计算方法模式[J]. 模式识别与人工智能，2013，26(5)：447-453

[10] 项海飞. 基于互信息粒度的相对约简的矩阵计算方法[J]. 西南师范大学学报：自然科学版，2014，39(3)：60-64

[11] Rizvi S J, Haritsa J R. Maintaining Data Privacy in Association Rule Mining[C]// Proc of the 28th Intl Conf on Very Large Data Bases (VLDB), 2002. Hong Kong, China, 2002: 682-693

[12] Agrawal S, Krishnan V, Haritsa J R. On Addressing Efficiency Concerns in Privacy-preserving Mining[C]// Proc of 9th Intl Conf on Database Systems for Advanced Applications (DASFAA), 2004. Jeju Island, Korea, 2004: 113-124

(上接第 189 页)

[9] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228-5235

[10] Chang C-C, Lin C-J. LIBSVM: A library for support vector machines[J]. ACM Trans. Intell. Syst. Technol., 2011, 2(3): 1-27

[11] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]. SIGKDD Explor. Newsl., 2009, 11(1): 10-18

[12] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on twitter[C]// Proceedings of the international conference on World Wide Web (WWW). 2011: 705-714

[13] Diggle P. A kernel method for smoothing point process data[J]. Applied Statistics, 1985, 34(2): 138-147