

基于美食互动社区的用户饮食行为模型研究

李越, 曹菡

(陕西师范大学 计算机科学学院, 陕西 西安 710062)

摘要:随着大数据、“互联网+”时代的到来,互联网美食互动社区的用户原创内容呈爆发式增长,从海量饮食数据中发现自己希望寻找的内容越来越不容易,同时该部分数据没有得到广泛的利用和深度的挖掘;传统的对于饮食行为的研究多采用问卷调查等形式,耗费了大量人力、物力、财力。针对以上问题,提出了基于LDA的用户饮食行为模型;利用LDA模型的思想,分析互联网美食互动社区的用户原创内容,根据困惑度确定主题数,构建用户饮食行为模型,进而可以计算用户饮食行为相似度,以此为美食社区用户进行好友和美食推荐提供模型基础,同时为饮食行为研究提供了一个新思路。以爬虫技术获取互联网美食互动社区上的用户原创内容作为数据集,通过实验验证了这种算法的可行性和有效性。

关键词:饮食行为;美食互动社区;用户模型;数据挖掘;LDA模型

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2016)12-0156-04

doi:10.3969/j.issn.1673-629X.2016.12.034

Research on User Eating Behavior Model Based on Food Interactive Community

LI Yue, CAO Han

(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: As the time for big data and "Internet+" era is coming, user generated content of Internet food interactive community is experiencing the explosive growth. It is becoming more and more difficult for users to find the content of interest. And this part of the data has not been widely used and deeply mined. Traditional eating behavior research normally uses questionnaire, which spends a lot of manpower, material and financial resources. To solve the above problem, it presents user eating behavior model based on LDA. In order to build this model, the ideas of LDA model is used to analyze user generated content of Internet food interactive community, determining the subject number of model according to the perplexity, then calculating the user similarity of eating behavior, which can provide a basis of recommending friends or food for community users. It also provides a new way of eating behavior research. The user generated content from a Internet food interactive community is collected as data set. The experiments verify the feasibility and effectiveness of this method.

Key words: eating behavior; food interactive community; user model; data mining; LDA model

0 引言

随着人民生活水平的不断提高,吃饱已经不能满足人们对饮食的需求,人们开始追求饮食的美味与健康。随着互联网技术和新的媒体形式的崛起,美食作为生活化互联网的一项服务,逐渐和网络社区结合成一种互联网美食经济产业链,由此催生的美食互动网站的设计和运营也变得越来越热门^[1]。美食互动社区的快速成长与发展是互联网持续向社会生活渗透的写照之一,为人们获取更多关于饮食方面的信息提供了支撑,为美食爱好者提供了一个在线交流平台。人们

通过美食互动社区发现、分享和交流美食。美食互动社区是典型的用户原创内容(User Generated Content, UGC)社区,其中80%的内容来自于用户。人们在网络中发布菜谱等这些线上行为一定程度上反映了用户线下的饮食行为习惯,这部分数据如果能得到充分的利用和挖掘,对于饮食行为干预^[2]、疾病预防和控制^[3]、食品推荐等问题的解决将起到很大的帮助。

传统的饮食行为研究方法通常是采用膳食调查^[4]的方法,通过问卷及24小时食物记录表^[5]的方式进行,耗费大量的人力物力不说,对于食物摄入量测量、

收稿日期:2016-01-20

修回日期:2016-05-18

网络出版时间:2016-10-24

基金项目:国家自然科学基金资助项目(41271387)

作者简介:李越(1991-),女,硕士研究生,研究方向为云计算、高性能计算、机器学习、数据挖掘;曹菡,教授,研究方向为并行计算、大数据处理、空间数据挖掘、智慧旅游。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20161024.1113.040.html>

食物成分多样性等复杂问题也得不到有效解决;第二类是对研究对象的调查,需要对研究对象进行跟踪记录,需要研究对象的主动参与。但上述方法均忽略了用户在互联网上留下的信息。

文中对美食社区数据进行统计分析,然后利用 LDA 模型构建用户饮食行为模型,以此模型为基础计算用户的相似度,为美食社区用户推荐和食品推荐提供模型基础。

1 LDA 模型的基本思想

LDA (Latent Dirichlet Allocation) 是目前应用最广泛的隐主题模型^[6],具有扎实的概率基础和可靠的扩展性,被广泛应用于文本建模的各个领域。LDA 是一个三层(文档-主题-词)贝叶斯模型,图 1 为 LDA 图模型表示。将文档表示成隐主题上的分布,而每个主题又表示成词的分布。

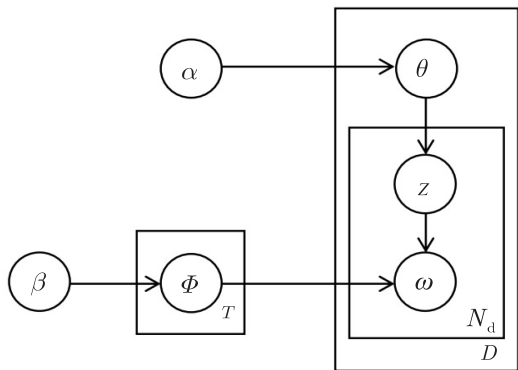


图 1 LDA 的图模型表示

其中,LDA 模型采用 Dirichlet 分布作为概率主题模型中多项分布的先验分布。 D 为整个文档集, N_d 为文档 d 的单词集, α 和 β 分别为文档-主题概率分布 θ 和主题-单词概率分布 Φ 的先验知识, T 为隐主题数。

2 基于 LDA 模型的用户饮食行为模型研究

2.1 基于 LDA 模型的用户饮食行为模型

文中借助于 LDA 模型的思想,构建用户饮食行为模型 (Author-Eating Behavior Model) 将原本的文档建模推广到用户饮食行为建模之上。假设数据集中的每个用户对应一个隐饮食行为的分布,而隐饮食行为则同样由菜谱属性词的分布表示。

使用 LDA 模型构建用户饮食行为模型时,需要将一个用户下的所有菜谱合并成一个文档进行饮食行为生成,从而得到用户饮食行为的概率多项分布,即用户的饮食行为模型。该模型将文档-主题-词的三层关系变成了用户-饮食行为-词的关系。

在用户层中,对用户集 $U = \{u_1, u_2, \dots, u_m\}$ 中每一个用户 u_i 及其所有菜谱得到一个词频向量 $f_{u_i} =$

$\langle tf_{i,1}, tf_{i,2}, \dots, tf_{i,v} \rangle$,从饮食行为层而言, u_i 可以被表示成向量 $\theta_{u_i} = \{p_{u_i,1}, p_{u_i,2}, \dots, p_{u_i,k}\}$ 。其中, $p_{u_i,k}$ 表示饮食行为 z 在用户 u_i 中的生成概率,用它来表示用户 u_i 对饮食行为 z 的依赖程度。因此,用户层构成了用户与饮食行为的生成关系,生成用户饮食行为模型,其

$$\text{矩阵为: } \begin{Bmatrix} p_{u_1,1} & p_{u_1,2} & \cdots & p_{u_1,k} \\ p_{u_2,1} & p_{u_2,2} & \cdots & p_{u_2,k} \\ \vdots & \vdots & \cdots & \vdots \\ p_{u_m,1} & p_{u_m,2} & \cdots & p_{u_m,k} \end{Bmatrix}。$$

同时,得到一个饮食行为-词的概率生成矩阵:

$$\begin{Bmatrix} p_{z_1,1} & p_{z_1,2} & \cdots & p_{z_1,v} \\ p_{z_2,1} & p_{z_2,2} & \cdots & p_{z_2,v} \\ \vdots & \vdots & \cdots & \vdots \\ p_{z_i,1} & p_{z_i,2} & \cdots & p_{z_i,v} \end{Bmatrix}。$$

其中, $p_{z_i,v}$ 为给定饮食行为 z 时生成词 w 的概率。

2.2 用户饮食行为相似性计算

相似用户具有相近的饮食行为。计算用户间的相似度,可以将其应用于美食社区进行用户和食品的推荐。

KL (Kullback Leibler) 散度,俗称 KL 距离^[7],常用来衡量两个概率分布的距离,其计算公式如下:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (1)$$

KL 散度是不对称的,即:

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P) \quad (2)$$

可以将其转换为对称的,如下所示:

$$D(P, Q) = [D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)]/2 \quad (3)$$

在基于 LDA 的用户饮食行为模型中,如用户主题矩阵所示,用户间的相似程度可以由各用户饮食行为分布之间的 KL 距离表示,用户相似度计算如下所示:

$$u_i = \frac{1}{D(U_i, U_j)} = \frac{2}{D_{KL}(U_i \parallel U_j) + D_{KL}(U_j \parallel U_i)} \quad (4)$$

其中, s_{ij} 为用户 u_i 和 u_j 的相似度; U_i 和 U_j 分别是他们的饮食行为概率分布。 s_{ij} 越大,表示两个用户越相似。

3 实验结果与分析

3.1 实验准备

应用爬虫技术,从某美食互动社区网站上随机获取 2014 年 4 月到 2015 年 3 月期间 6 834 篇美食博客数据,数据概要如表 1 所示。

通过统计分析发现:

(1) 工艺为“炒”的菜谱最多,占总数的 24.5%,其次为“煮”,占 16.4%，“拌”占 12%。在中国,大部分家庭蔬菜烹饪以炒菜为主^[8],数据统计符合中国人的

传统饮食习惯。

(2)最多食类主料依次为猪肉、鸡蛋、面粉、胡萝卜、土豆、虾、大米、西红柿、豆腐、木耳、青椒、洋葱、牛奶、低筋面粉、香菇。均为日常生活中常见食材,便于

获取,烹饪简单。

(3)“两人份”菜谱占 49.2% ,“三人份”菜谱占 25% 。与中国家庭结构吻合。

表 1 数据概要

用户 名	菜谱 名	工艺	难度	人数	口味	准备时间	烹饪时间	主料	辅料	步骤
.....	炒、蒸、煮、炖、拌、烧、煎、炸、烘焙、微波、烤、煲、焖、冻、烙、砂锅、腌、卤、酱、烩、扒、爆、炆、熘、熏、余、拔丝、榨汁、灼、泡、腊、糖蘸、干锅、焗、干煸、煨、刺身、其他工艺	新手尝试、 初级入门、 初中水平、 中级掌勺、 高级厨师、 厨神级	1 人份、 2 人份、 3 人份、 4 人份、 5 人份、 未知	家常味、香辣味、咸鲜味、甜味、酸甜味、酸辣味、麻辣味、酱香味、奶香味、蒜香味、鱼香味、葱香味、果味、五香味、咖喱味、椒麻味、茄汁味、酸味、苦香味、姜汁味、怪味、芥末味、红油味、豆瓣味、麻酱味、黑椒味、糊辣味、其他	5 分钟、10 分钟、15 分钟、30 分钟、60 分钟、90 分钟、2 小时、数小时、一天、数天	<5 分钟、<10 分钟、<15 分钟、<30 分钟、<60 分钟、<90 分钟、<2 小时、<数小时、<一天、<数天

(4)准备时间在“15 分钟”以下的菜谱占 78.9% ,烹饪时间在“30 分钟”以下的占菜谱数的 69.3% 。说明人们倾向于简单易烹饪的食物。

(5)口味方面:“家常味”占 36.5% ,“咸鲜味”占 19.9% ,“甜味”占 15.1% 。

以上统计分析结果均与实际相符合,说明了网络数据的真实性、实用性,具有研究价值。

3.2 困惑度

困惑度^[9]是用来评价主题模型的一个重要指标,主题模型用概率分布来描述一个文本的生成过程,因此理所当然地会想到用熵的概念来评判主题模型是否有效。直观的解释即为:若词表中所有的词都具有统一的概率分布,即每个词出现的概率都是一样的,这种情况下是最难预测的,而由熵的概念知此时的熵最大。而概率分布越不均匀,熵值越小。

文中应用 LDA 模型构建的用户饮食行为模型属于主题模型的一种,故也选用困惑度作为衡量算法的标准。该模型中困惑度的公式如下:

$$\text{Perplexity}(D) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(W_d)}{\sum_{d=1}^M N_d} \right\} \quad (5)$$

$$p(W_d) = \prod_{n=1}^{N_d} \prod_{k=1}^K p(w_n | z_n = k) \cdot p(z_n = k | d) \quad (6)$$

$$\log p(W_d) = \sum_{n=1}^{N_d} \log \left(\sum_{k=1}^K \theta_{dk} \cdot \varphi_{k,w_n} \right) \quad (7)$$

其中, M 为测试集 D 中的用户数; $p(W_d)$ 为用户 d 的菜谱词向量; N_d 为该词向量的长度; K 为饮食行为数; $p(z_n = k | d)$ 为用户 d 产生饮食行为 z 的概率; $p(w_n | z_n = k)$ 为饮食行为 z 生成词 w 的概率; θ 为饮食行为的概率分布矩阵(见 2.1 节); φ 为词的概率分布矩阵(见 2.1 节)。

LDA 模型数据解过程使用基于吉布斯(Gibbs)抽

样的参数估计方法^[10-11],模型参数根据文献[12-15]选取经验值。其中, $\alpha = 50/K$ (K 为主题数,对应文中用户饮食行为模型中的隐饮食行为数), $\beta = 0.01$ 。根据困惑度的结果确定最佳的 K 值。首先,尝试设置 K 为 10,20,⋯,110 时的情况,如图 2(a)所示。模型的困惑度随着 K 的增大而减小,当 K 为 40 时困惑度最小,模型的效果最好。随着 K 不断增大,困惑度也随之增大。因此认为 K 的最优值在 40 附近。为进一步确定 K 的值,以 1 为间隔,选取 K 为 30~50 时计算困惑度,如图 2(b)所示。最终确定文中构建用户饮食模型时的 K 为 47。

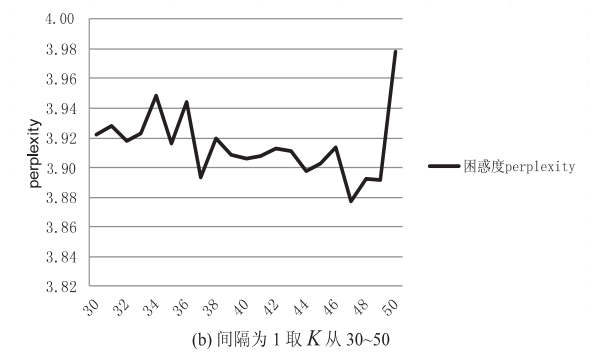
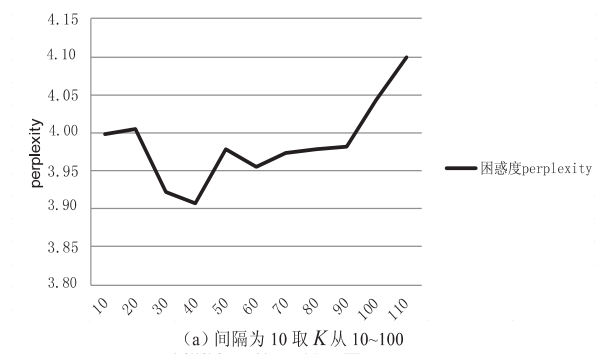


图 2 不同主题数下的困惑度

3.3 用户相似度

对采集到的数据进行随机筛选,以 30 个用户为例,应用饮食行为模型分析用户间的相似度,设置饮食

行为 $K=47$, 得到相似度矩阵。随机抽取一位用户, 列出与其相似度最高的十位用户, 如表 1 所示。可根据

用户之间的相似关系提供食品推荐服务、群体饮食行为研究等。

表 2 与用户 1 相似度最高的十位用户

s_{ij}	user25	user7	user28	user29	user16	user5	user23	user4	user24	user14
user1	1.42	1.29	1.11	1.05	1	0.85	0.84	0.78	0.78	0.77

4 结束语

针对美食互动社区中的 UCG 数据, 结合 LDA 模型的文档-主题-词分层模型的特点, 用 UCG 数据来代表用户, 进而提出了用户-饮食行为-词的用户饮食行为模型, 为数据挖掘在饮食行为方面的研究提供了一个新思路。今后的研究工作可结合更多的社交网络特征, 通过数据挖掘, 为解决饮食行为干预、疾病预防和控制、食品推荐等问题提供更大的帮助。

参考文献:

[1] 毛茅, 王洋, 赵好婕, 等. 基于社交网络的美食互动网站设计与评估[C]//第七届和谐人机环境联合学术会议 (HHME2011) 论文集. 出版地不详; 出版者不详, 2011.

[2] 杨正雄, 赵文华, 陈君石. 饮食行为干预的研究进展[J]. 中国学校卫生, 2008, 29(6): 573-576.

[3] 贡浩凌, 戴莉敏, 刘媛, 等. 医院-社区-家庭护理干预模式对 2 型糖尿病患者饮食控制的效果[J]. 中华护理杂志, 2014, 49(4): 399-403.

[4] 张雅楠, 丁虹, 杜玉萍. 回顾性膳食调查辅助工具的应用现状与评价方法[J]. 职业与健康, 2015(9): 1294-1296.

[5] 安宜沛. 慢性心衰患者膳食现状调查及中医药膳调养研究[D]. 广州: 广州中医药大学, 2015.

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

[7] 孙昌年, 郑诚, 夏青松. 基于 LDA 的中文文本相似度计算[J]. 计算机技术与发展, 2013, 23(1): 217-220.

[8] 曾利明. 中国民众存在五大饮食“误区”[N]. 光明日报, 2004-11-26.

[9] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1-2): 177-196.

[10] 张斌, 张引, 高克宁, 等. 融合关系与内容分析的社会标签推荐[J]. 软件学报, 2012, 23(3): 476-488.

[11] Griffiths T, Steyvers M. Probabilistic topic models[M]//Latent semantic analysis. Hillsdale, NJ: Laurence Erlbaum, 2006.

[12] Asuncion A, Welling M, Smyth P, et al. On smoothing and inference for topic models[C]//Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. [s. l.]: AUAI Press, 2009: 27-34.

[13] 石晶, 胡明, 石鑫, 等. 基于 LDA 模型的文本分割[J]. 计算机学报, 2008, 31(10): 1865-1873.

[14] 刘振鹿, 王大玲, 冯时, 等. 一种基于 LDA 的潜在语义区划分及 Web 文档聚类算法[J]. 中文信息学报, 2011, 25(1): 60-65.

[15] 李文峰. 基于主题模型的用户建模研究[D]. 北京: 北京邮电大学, 2013.

(上接第 155 页)

[J]. 河南科技, 2013(6): 3-5.

[4] 侯翠琴, 焦李成, 张文革. 一种压缩稀疏用户评分矩阵的协同过滤算法[J]. 西安电子科技大学学报, 2009, 36(4): 614-618.

[5] Ar Y, Bostanci E. A genetic algorithm solution to the collaborative filtering problem[J]. Expert Systems with Applications, 2016, 61: 122-128.

[6] 李晓城, 张增杰, 夏勇明, 等. 基于 web 数据挖掘的健康餐饮分析推荐系统的设计[J]. 微型电脑应用, 2011, 27(1): 44-46.

[7] 付德坤. 基于模糊决策的中医饮食推荐建模及嵌入式系统实现[D]. 成都: 电子科技大学, 2013.

[8] 黄洋. 基于聚类和项目类别偏好的协同过滤推荐算法研究[D]. 杭州: 浙江理工大学, 2014.

[9] Abdelwahab A, Sekiya H, Matsuba I. Collaborative filtering based on an iterative prediction method to alleviate the sparsity

problem[C]//Proceedings of the 11th international conference on information integration and web-based applications & services. [s. l.]: ACM, 2009.

[10] Tino P. Bifurcation structure of equilibria of iterated Softmax[J]. Chaos, Solitons & Fractals, 2009, 41(4): 1804-1816.

[11] 付鹏, 姚建刚, 龚磊. 利用红外特征和 Softmax 回归识别绝缘子污秽等级[J]. 计算机工程与应用, 2015, 51(13): 181-185.

[12] 汪海波, 陈雁翔, 李艳秋. 基于主成分分析和 Softmax 回归模型的人脸识别方法[J]. 合肥工业大学学报: 自然科学版, 2015, 38(6): 759-763.

[13] 王晟. 基于 Softmax 回归的电力仪表分类[J]. 研究与开发, 2014(6): 25-28.

[14] Majid A, Chen Ling, Chen Gencai. A context-aware personalized travel recommendation system based on geotagged social media data mining[J]. International Journal of Geographical Information Science, 2013, 27(4): 662-663.