



文章编号:1001-9081(2017)11-3201-06

DOI:10.11772/j.issn.1001-9081.2017.11.3201

融合异质网络与主题模型的方面分预测

吉余岗^{1,2}, 李依桐^{1,2}, 石川^{1,2*}

(1. 北京邮电大学 计算机学院, 北京 100876; 2. 智能通信软件与多媒体北京市重点实验室(北京邮电大学), 北京 100876)

(* 通信作者电子邮箱 shichuan@bupt.edu.cn)

摘要:针对传统方面分预测模型只考虑内容信息而缺乏对评论网络结构的分析,提出了融合异质信息网络和主题模型构建方面分预测算法(HINToAsp)。首先,从意见短语角度构建了评论主题挖掘模型(Phrase-PLSA),有效整合评论信息和评分信息进行方面主题挖掘;进而,考虑用户、评论和商品之间的结构信息,提出了在“用户-评论-商品”异质信息网络上的主题传播模型模型,用于刻画用户特性、商品属性;最后,基于随机游走框架有效整合内容信息和结构信息,进行精准的方面分预测。通过在大众点评(Dianping)和 TripAdvisor 数据集上和四元组 PLSA(QPLSA)、高斯分布的情绪评估(GRAOS)模型及情绪均衡主题模型(SATM)的准确度对比实验,证明了 HINToAsp 算法的有效性,可以更好地用于商品的推荐系统。

关键词:方面分预测;异质信息网络;主题模型;结构信息;推荐系统

中图分类号:TP391 文献标志码:A

Aspect rating prediction based on heterogeneous network and topic model

JI Yugang^{1,2}, LI Yitong^{1,2}, SHI Chuan^{1,2*}

(1. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia

(Beijing University of Posts and Telecommunications), Beijing 100876, China)

Abstract: Concerning the problem that traditional aspect rating prediction methods just pay attention to textual information while ignoring the structural information in the review network, a novel Aspect rating prediction method based on Heterogeneous Information Network and Topic model (HINToAsp) was proposed for effectively integrating textual information and structural information. Firstly, a new review topic model of opinion phrases called Phrase-PLSA (Phrase-based Probabilistic Latent Semantic Analysis) was put forward to integrate textual information of reviews and ratings for mining aspect topics. And then, considering the rich structural information among users, reviews, and items, a topic propagation model was designed by the aid of constructing “User-Review-Item” heterogeneous information network. Finally, a random walk framework was used to combine textual information and structural information effectively, which insured an accurate aspect rating prediction. The experimental results on both Dianping corpora and TripAdvisor corpora demonstrate that HINToAsp is more effective than recent methods like the Quad-tuples PLSA (QPLSA) model, the Gaussian distribution for RAting Over Sentiments (GRAOS) model and the Sentiment-Aligned Topic Model (SATM), and has better performance on recommendation system.

Key words: aspect rating prediction; Heterogeneous Information Network (HIN); topic model; structural information; recommendation system

0 引言

近年来,电商平台和团购网站蓬勃发展,逐渐改变了人们的生活和消费方式。在这些平台上,用户可以通过打分和撰写评论来对商品的各方面质量进行评价,商品的评价信息会极大影响后续消费者的流量^[1]。为了从这些大量的评价信息中快速总结出商品各方面的质量优劣进而用于商品推荐,人们开始关注方面分预测研究。

方面分预测的主要任务是预测用户对商品各方面的评分。为了实现有效的方面分预测,通常需要选择有效的文本表示模型来表征文字评论信息。而主题模型因其低维密实和解释性强等原因,受到研究者的青睐^[2-3]。

传统的主题模型,如概率潜在语义分析(Probabilistic Latent Analysis, PLSA)^[4]和潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)^[5]等,常用于分析单词的主题分布,因此,这些模型用于挖掘评论主题时,忽视评论中意见短语强烈的主题指向。针对评论信息的特性,Lu 等^[6]提出一种改进的 PLSA 模型来识别评论短语的主题。

当前,方面分预测算法多从内容信息角度来提取特征,如总分和评论的主题分布,却忽视用户和商品间的关联特征。而用户对不同商品的不同评分和评论,实际上构建出一个典型的异质信息网络(Heterogeneous Information Network, HIN)^[7],而 HIN 中包含了丰富的结构特征,广泛用于解决推荐系统问题^[8]。

收稿日期:2017-05-11;修回日期:2017-05-31。

基金项目:国家自然科学基金资助项目(61375058);国家973计划项目(2013cb329606);北京市教育委员会共建项目。

作者简介:吉余岗(1993—),男,江苏泰州人,博士研究生,CCF会员,主要研究方向:数据挖掘、机器学习;李依桐(1992—),女,北京人,硕士,主要研究方向:数据挖掘、机器学习;石川(1978—),男,北京人,教授,博士,CCF会员,主要研究方向:数据挖掘、机器学习、演化计算。



鉴于前人的研究,本文考虑内容信息和结构信息,提出融合异质信息网络和主题模型的方面分预测算法 HINToAsp。首先,从评论短语和总分角度构建了一种 Phrase-PLSA 模型,用于识别短语的主题;然后,提出了一种基于评论行为的异质信息网络,通过评论的主题分布传递给用户和商品来刻画用户特性和商品属性;最后,在随机游走框架下将内容信息和结构信息有效整合后预测方面分。

本文主要贡献如下:

- 1) 基于用户对商品的评论数据,构建了评论行为的异质信息网络,有效刻画用户特性和商品属性;
- 2) 分别基于 Phrase-PLSA 和 HIN 来发现评论数据的内容信息和结构信息,并提出了一种随机游走框架将两者有效整合;
- 3) 在中文和英文评论数据集上不同规模的方面分预测实验,有效证明了所提算法的有效性和泛化性能。

1 相关工作

结合评论和评分信息成为解决方面分预测的关键技术。

Zheng 等^[9]提出一种评价表达模式的 LDA (Appraisal-Expression-Patterns-based LDA, AEP-LDA) 模型,自动从评论中提取方面词;Wang 等^[10]提出潜在方面评分分析模型 (Latent Aspect Rating Analysis Model, LARAM) 算法,从方面级角度分析评论中的观点,并以此来预测用户对各方面的评分;文献 [11] 提出通过外部知识、总分分布以及情感词语词典等同步提取方面主题及对应评分;Li 等^[12]提出了一种考虑用户评分偏好影响的 PLSA 模型。但这些模型普遍只考虑了文本内容信息,忽视了评论网络中丰富的结构信息。

异质信息网络常用于建模社会媒体系统中不同类型的对象和对象间繁杂的交互关系。许多推荐方法通过 HIN 来整合各类信息:Shi 等^[13]提出了异质网络上的电影推荐系统 (Heterogeneous network Recommendation, HeteRecom),通过元路径包含的语义信息计算电影之间的相似性;Yu 等^[14]基于元路径隐藏特征建模用户和商品之间的内在联系,分别从全局及个性化角度设计推荐模型;Sun 等^[15]提出了在科研学术网络上主题建模,并构建科研学术异质网络用于挖掘论文作者相似性;张邦佐等^[16]提出融合异质信息网络和矩阵分解进行总分预测。这些研究表明,在异质信息网络上的结构信息是可靠合理的。

2 PLSA 和 HIN

2.1 PLSA 模型

PLSA 模型通过期望最大化 (Expectation Maximization, EM) 算法学习相关参数。图 1 是 PLSA 的概率图模型。

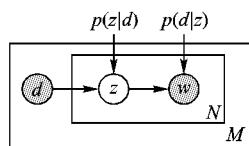


图 1 PLSA 模型概率图

Fig. 1 Probabilistic graph of PLSA model

图 1 中, d 表示一篇文档, z 表示隐含主题或方面, w 表示文档中的词语。 $p(d_i)$ 表示文档 d_i 出现的概率, $p(z_k | d_i)$ 表示文档 d_i 中出现主题 z 为 k 的概率,是一个多项分布。 $p(w_j | z_k)$

表示主题 k 下出现词语 w_j 的概率,也是一个多项分布。图 1 中 d 、 w 为可观测变量,主题 z 为隐藏变量,则可观测数据 (d_i, w_j) 的联合概率分布如下:

$$p(d_i, w_j) = p(d_i) \sum_{k=1}^K p(w_j | z_k) p(z_k | d_i) \quad (1)$$

其中: $i \in \{1, 2, \dots, M\}$, M 为文档集大小, $j \in \{1, 2, \dots, N\}$, N 为词的总数, $k \in \{1, 2, \dots, K\}$, K 表示主题总数。通过 EM 算法来学习式(1)中的参数 $p(w_j | z_k)$ 和 $p(z_k | d_i)$ 。

2.2 HIN 概念

异质信息网络是一种以有向图为数据结构的特殊的信息网络,可以包含多类型对象以及多类型的边。

定义 1 异质信息网络。给定一个模式 (A, R) , 其中 A 表示实体集, R 表示关系集。信息网络被定义为有向图 $G = (V, E)$, 其中对象类型映射函数为 $\Phi: V \rightarrow A$, 关系类型映射函数为 $\Psi: E \rightarrow R$ 。每个对象 $v \in V$ 属于某一特定的对象类型 $\Phi(v) \in A$, 每条边 $e \in E$ 属于某一特定的关系类型 $\Psi(e) \in R$ 。当对象种类 $|A| > 1$ 或关系种类 $|R| > 1$ 时, 此网络即为异质信息网络。

定义 2 元路径。元路径 P 是在模式 (A, R) 上的路径, 表示为 $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_L} A_{L+1}$, 简写为 $A_1 A_2 \dots A_{L+1}$, 定义 A_1 和 A_{L+1} 中存在复合关系 $R = R_1 \circ R_2 \circ \dots \circ R_L$, 其中 \circ 表示关系间的复合操作符, L 表示复合关系数。

异质信息网络可以有效融合更多的结构信息、包含更丰富的语义,是数据挖掘领域的一个新的方向,异质信息网络用于推荐时,可以更加细致地描述用户和商品间的关系。

3 HINToAsp 算法

方面分预测的主要挑战是评论的文本建模以及和评分的结合。本文提出一种基于异质信息网络和主题模型的方面分预测算法 (Aspect rating prediction method based on Heterogeneous Information Network and Topic model, HINToAsp), 分别从内容信息和结构信息两个角度构建了 Phrase-PLSA 和 Review HIN 模型。通过 Phrase-LDA, 以短语为单位构建主题模型, 挖掘出短语的主题分布;进而通过 Review HIN 充分考虑用户和商品间的链接信息, 有效刻画用户行为特性和商品属性;通过随机游走框架将两部分结合一起。模型结构如图 2 所示。

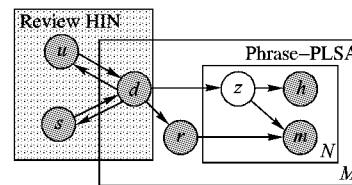


图 2 HINToAsp 模型

Fig. 2 HINToAsp model

其中, 阴影框表示 Review HIN 的网络模式, 阴影框中箭头表示链路连接关系; 右侧为 Phrase-PLSA 概率图模型。涉及的概念定义如下。

用户 (User): 用户 u 表示用户集合 U 中的一人。

物品 (Item): 物品 s 表示物品集合 S 中的一个商品 (如大众点评数据中的餐馆)。

评论 (Review): 评论 d 表示用户 u 对物品 s 的文本评价信



息。

短语(Phrase):由从评论 d 中抽取的一对词语 $\langle h, m \rangle$ 组成, h 表示先行词, m 表示修饰词。

先行词(Head Term):先行词 h 描述方面信息。

修饰词(Modifier Term):修饰词 m 描述情感信息。

总评分(Overall Rating):每条评论 d 对应的总评分 r , 通常为 1 ~ 5 的整数评分。

方面(Aspect):方面 z 表示物品 s 的一个属性或方面。

方面评分(Aspect Rating):方面评分 a_z 表示物品 s 在 z 方面的打分。

3.1 Phrase-PLSA 模型

本文提出一种改进的 Phrase-PLSA 模型, 用于融合评论和评分等内容信息进行主题挖掘和方面分预测, 图 3 为对应的概率图。

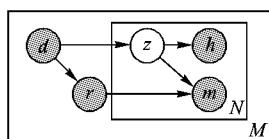


图 3 改进的 Phrase-PLSA 模型概率图

Fig. 3 Probabilistic graph of improved Phrase-PLSA model

Phrase-PLSA 采用 EM 算法推导参数迭代计算公式, 详细推导过程如下。

E 步中, 需要构造下界函数 \mathcal{L}_0 , 计算公式如下:

$$\mathcal{L}_0 = \sum_{k=1}^K q(z_k) \lg \left\{ \frac{p(h, m, r, d, z_k | A)}{q(z_k)} \right\} = \mathcal{L} - const \quad (2)$$

其中: $\mathcal{L} = \sum_{k=1}^K q(z_k) \lg q(z_k) \lg p(h, m, r, d, z_k | A); const = \sum_{k=1}^K q(z_k) \lg q(z_k); A$ 由 $p(m | r, z_k), p(h | z_k), p(z_k | d), p(r | z_k)$ 及 $p(d)$ 等所有参数组成。 $q(z_k)$ 表示隐含变量的后验概率, 如式(3)所示:

$$q(z_k) = p(z_k | h, m, r, d; A^{old}) = \frac{p(m | r, z_k) p(h | z_k) p(z_k | d) p(r | d) p(d)}{\sum_{i=1}^K p(m | r, z_i) p(h | z_i) p(z_i | d) p(r | d) p(d)} \quad (3)$$

因此, 每次迭代过程中, 式(2)中的 $const$ 只与上一轮的结果有关, 只需最大化 \mathcal{L} 即可:

$$\mathcal{L} = \sum_{i=1}^M \sum_{j_h=1}^{N_h} \sum_{j_m=1}^{N_m} \sum_{k=1}^K \sum_{s=1}^R n(h_{j_h}, m_{j_m}, r_s, d_i, z_k) q(z_k) \cdot \lg p(h_{j_h}, m_{j_m}, r_s, d_i, z_k | A) \quad (4)$$

其中: $p(h_{j_h}, m_{j_m}, r_s, d_i, z_k | A) = p(m_{j_m} | r_s, z_k) p(h_{j_h} | z_k) p(z_k | d_i) p(r_s | d_i) p(d_i)$, N_h 为先行词总数, N_m 表示修饰词总数。

M 步中, 最大化下界函数 \mathcal{L}_0 , 基于以上推导, 即最优化 \mathcal{L} 使用拉格朗日乘子法来最大化 \mathcal{L} 并计算参数。对于参数

$p(m_{j_m} | r_s, z_k)$, 存在 $\sum_{j_m=1}^{N_m} p(m_{j_m} | r_s, z_k) = 1$ 。应用拉格朗日乘子法, 得到关于 $p(m_{j_m} | r_s, z_k)$ 的函数:

$$\frac{\partial \left[\mathcal{L}_{p(m_{j_m} | r_s, z_k)} + \lambda \left(\sum_{j_m=1}^{N_m} p(m_{j_m} | r_s, z_k) - 1 \right) \right]}{\partial p(m_{j_m} | r_s, z_k)} = 0 \quad (5)$$

计算得到:

$$p(m_{j_m} | r_s, z_k) \propto n(h_{j_h}, m_{j_m}, r_s, d_i) p(z_k | h_{j_h}, m_{j_m}, r_s, d_i; A^{old}) \quad (6)$$

因此 $p(m_{j_m} | r_s, z_k)$ 的更新函数为:

$$p(m_{j_m} | r_s, z_k) = \frac{\sum_{h, d} n(h, m_{j_m}, r_s, d) p(z_k | h, m_{j_m}, r_s, d; A^{old})}{\sum_{h, m' d} n(h, m', r_s, d) p(z_k | h, m', r_s, d; A^{old})} \quad (7)$$

同理, 其他参数的更新函数为:

$$p(h_{j_h} | z_k) = \frac{\sum_{m, r, d} n(h_{j_h}, m, r, d) p(z_k | h_{j_h}, m, r, d; A^{old})}{\sum_{h', m, r, d} n(h', m, r, d) p(z_k | h', m, r, d; A^{old})} \quad (8)$$

$$p(z_k | d_i) = \frac{\sum_{h, m, r} n(h, m, r, d_i) p(z_k | h, m, r, d_i; A^{old})}{\sum_{h, m, r, z'} n(h, m, r, d_i) p(z' | h, m, r, d_i; A^{old})} \quad (9)$$

$$p(r_s | d_i) = \frac{\sum_{h, m, r} n(h, m, r, d_i) p(z | h, m, r, d_i; A^{old})}{\sum_{h, m, r, z} n(h, m, r, d_i) p(z | h, m, r, d_i; A^{old})} \quad (10)$$

$$p(d_i) = \frac{\sum_{h, m, r, z} n(h, m, r, d_i) p(z | h, m, r, d_i; A^{old})}{\sum_{h, m, r, d'} n(h, m, r, d') p(z | h, m, r, d'; A^{old})} \quad (11)$$

3.2 评论异质信息网络

在购物或消费过程中, 不同用户对不同商品撰写对应的评论文本, 这种行为构成了一种评论网络, 如图 4(a)所示。本文提出构建基于评论的异质网络, 其模式如图 4(b)所示。网络中有用户(U)、商品(S)、评论(D)等三种类型的节点, 同时包含了多种元路径及其蕴含的物理意义, 如 $u_1 d_1 s_1$ 表示用户 u_1 对商品 s_1 撰写评论 d_1 。

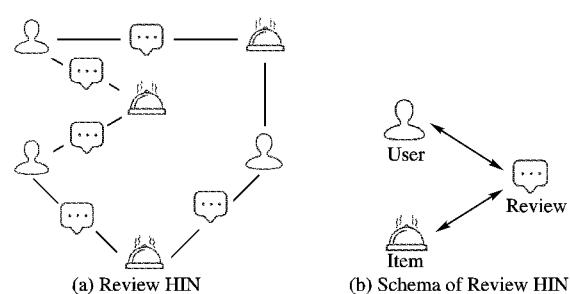


图 4 评论异质信息网络及其模式

Fig. 4 Structure of Review HIN and its' schema

主题在评论 D 和与其相关的用户 U 和商品 S 间传播。给定一条评论的主题分布 $p(z_k | d_i)$, 一个用户 u 的主题分布按式(12)计算:

$$p(z_k | u) = \sum_{d_i \in D_u} p(z_k | d_i) p(d_i | u) \quad (12)$$

其中: D_u 表示由 u 撰写的评论集合。相似地, 一个商品 s 的主题分布的计算公式如下:

$$p(z_k | s) = \sum_{d_j \in D_s} p(z_k | d_j) p(d_j | s) \quad (13)$$

另一方面, 主题分布也可以从用户 U 和商品 S 传播到评论 D 中。根据通过主题模型估算出的评论的内在主题分布,



提出如下主题传播算法:

$$p(z_k | d_i) = \xi p(z_k | d_i) + \frac{(1 - \xi)}{2} [p(z_k | u) + p(z_k | s)] \quad (14)$$

其中: d_i 是用户 u 对商品 s 的评价。 ξ 表示主题传播过程中, 传播偏好参数, 用于调节 Phrase-PLSA 中挖掘的主题分布和 Review HIN 上传播的主题分布的权重影响: ξ 为 0 表示算法仅考虑 Review HIN 部分; ξ 为 1 表示算法仅考虑 Phrase-PLSA 部分, 称之为 HINToAsp\`s。

3.3 方面识别和方面分预测

为验证模型有效性, 需要将预测的方面评分与真实方面的评分对比。由于预测方面应当与语料库中的要求的方面相对应, 因此在构建模型时需要给每个方面预设部分先验词语, 如后文 4.1 节。

在实验中, 方面 z 加入先验知识, 计算公式为:

$$\begin{aligned} p(h_{jm} | z_k) &= \\ &\frac{\sum_{m,r,d} n(h_{jm}, m, r, d) p(z | h_{jm}, m, r, d; \Lambda^{\text{old}}) + \tau(h_{jm}, z_k)}{\sum_{h',m,r,d} n(h', m, r, d) p(z | h', m, r, d; \Lambda^{\text{old}}) + \tau(h', z_k)} \end{aligned} \quad (15)$$

其中: $\tau(h_{jm}, z_k)$ 表示词语的先验信息, 当 h_{jm} 的主题为 z_k 时, $\tau(h_{jm}, z_k) = 1$, 否则 $\tau(h_{jm}, z_k) = 0$ 。

方面识别 根据从模型中学习的参数, 基于式(16)计算出对应的 phrase 属于的方面。

$$\begin{aligned} z_{\langle h, m \rangle} &= \arg \max_z \sum_r p(z, r | h, m) = \\ &\arg \max_z \frac{\sum_{r,d,z} p(h, m, r, d, z)}{\sum_{r,d,z'} p(h, m, r, d, z')} \end{aligned} \quad (16)$$

方面分预测 给定若干短语 $\{\langle h, m \rangle\}$ 所描述的实体 e 时, 预测方面 z 上的得分 a_z 。预测公式如下:

$$a_z = \frac{\sum_{\langle h, m \rangle \in e} \sum_r r \cdot p(z, r | h, m)}{\sum_{\langle h, m \rangle \in e} \sum_r p(z, r | h, m)} \quad (17)$$

3.4 统一模型

融合 HIN 和 Phrase-PLSA 的 HINToAsp 算法的具体步骤如下。

输入 评论集 D , 对应评论短语集 $\{\langle h, m \rangle\}$, 集合 R , 用户集 U , 商品集 S , 评论短语先验信息;

输出 每个 phrase 属于的主题及对应评分。

1) 随机初始化 $p(m_{jm} | r_s, z_k), p(h_{jh} | z_k), p(z_k | d_i), p(d_i), p(r_s | d_i)$ 依据式(15)更新 $p(h_{jm} | z_k)$ 。

2) E 步: 计算给定参数 $p(m_{jm} | r_s, z_k), p(h_{jh} | z_k), p(z_k | d_i), p(r_s | d_i), p(d_i)$ 时隐藏变量的后验概率, 即 $p(h_{jh}, m_{jm}, r_s, d_i, z_k | \Lambda)$ 。

3) M 步: 最大化下界函数 \mathcal{L}_0 , 根据式(7)~(11)更新参数 $p(m_{jm} | r_s, z_k), p(h_{jh} | z_k), p(z_k | d_i), p(r_s | d_i), p(d_i)$ 。

4) 返回步骤 2) 继续迭代, 直至收敛结束迭代。

5) 依据式(16)计算得到在 Phrase-PLSA 上挖掘的评论短语主题。

6) 根据式(12)~(13)将评论集合的主题传递给与其相关的用户集 U 和商品集 S 。

7) 根据式(14)将用户 U 和商品(如餐馆) S 的主题分布传播到相关的评论集合 D 。

8) 返回步骤 5) 继续迭代直至收敛结束迭代。

9) 固定 $p(z_k | d_i)$, 重复步骤 2)~4) 的 EM 迭代, 直至收敛结束迭代。

10) 依据式(16), (17)计算融合 Phrase-PLSA 和 HIN 信息, 得到每条评论短语的主题及对应评分。

4 实验与分析

本章在大众点评(Dianping)和 TripAdvisor 的数据集上进行了不同规模的实验, 验证了 HINToAsp 的有效性和泛化性能。

4.1 数据预处理及参数设置

实验选取数据集为大众点评和 TripAdvisor 应用上采集的数据集。大众点评是一个集合餐饮娱乐等商家的中文社交媒体平台, 消费者可以在上面对商家的“口味”“服务”“环境”等方面评分, 并撰写评论。与“大众点评”相似, TripAdvisor 上用户的评价包括了总分, 英文评论以及在“价值”(Value)“服务”(Service)和“食物”(Food)方面上评分。数据集的统计信息如表 1 所示。

表 1 两个数据集上的统计信息

Tab. 1 Statistical information of two datasets

数据集	Users	Items	Reviews	Phrases
Dianping	14 519	1 097	216 291	696 608
TripAdvisor	192 108	5 579	437 088	4 562 247

数据预处理 主要是从评价中抽取短语, 由于两个数据集是不同语言的, 所以需要不同的预处理过程。TripAdvisor 数据集的预处理过程与文献[1]相似, 过程为: 1) 利用 POS (Part-Of-Speech) Tagging 标注词性; 2) 根据词性标注及文献[1]中的规则提取短语; 3) 采用 Porter Stemmer 进行词根还原。而处理大众点评数据集时, 不需要词根还原, 但在标注词性之前需要分词。本文采用 Word Segmenter 中文分词工具。

先验信息 本文选取先验评论短语见表 2。

表 2 两个数据集上的先验词语

Tab. 2 Prior terms of two datasets

数据集	方面	先验词
Dianping	Taste	taste, flavor, dish, dishes
	Service	serving, attitude, waitress, service
	Environment	environment, location, room, decoration
TripAdvisor	Value	value, price, quality, worth
	Service	service, attitude, waiter, waitress
	Food	food, taste, dish, dinner

4.2 评价指标及对比实验

实验采用均方根误差(Root Mean Square Error, RMSE)和皮尔逊相关系数(Pearson Correlation Coefficient, PCC)两个评价指标来评价模型有效性。其中: RMSE 用于衡量预测值和真实值之间的误差, 值越小则算法效果越好; PCC 用于衡量集合数据之间的线性关系, 比较预测值和实际值是否有相同的趋势变化, 值越接近于 1 则相关性越强。RMSE 和 PCC 的计算公式如下:

$$RMSE = \sqrt{\frac{\sum_{i=1}^M \sum_{k=1}^K (\hat{a}_{d_i, z_k} - a_{d_i, z_k})^2}{M \times K}} \quad (18)$$



$$PCC = \frac{M \times K \sum_{i=1}^M \sum_{k=1}^K \hat{a}_{d_i, z_k} a_{d_i, z_k} - \sum_{i=1}^M \sum_{k=1}^K \hat{a}_{d_i, z_k} \bar{a}_{d_i, z_k}}{\sqrt{M \times K \sum_{i=1}^M \sum_{k=1}^K (\hat{a}_{d_i, z_k})^2 - \sum_{i=1}^M \sum_{k=1}^K (\hat{a}_{d_i, z_k})^2} \times \sqrt{M \times K \sum_{i=1}^M \sum_{k=1}^K (a_{d_i, z_k})^2 - \sum_{i=1}^M \sum_{k=1}^K (a_{d_i, z_k})^2}} \quad (19)$$

本文实验和三个方面分预测的代表性方法四元组 PLSA (Quad-tuples PLSA, QPLSA)^[17]、高斯分布的情绪评估 (Gaussian distribution for Rating Over Sentiments, GRAOS) 模型^[18]及情绪均衡主题模型 (Sentiment-Aligned Topic Model, SATM)^[11]进行了准确度效果对比;此外,还对比了只使用主题模型而忽略异质网络的HINToAsp的模型,称之为

HINToAsp's。

QPLSA 提出了一个四元组概率隐藏语义分析模型,四元组是指先行词、修饰词、实体和评分; GRAOS 是一个半监督的 LDA 模型。模型从带有总评分的训练数据中挖掘出带有打分的方面信息,用于分析未被打分的数据的总评分; SATM 提出了情感排列主题模型,引入了情感词典和总评分分布这两类额外的信息用于评分预测。

4.3 准确性实验

采用 RMSE 评价模型方面分预测的准确性,设定主题个数 $K=3$,实验在规模分别为 25%、50%、75%、100% 数据集上进行实验,实验结果见表 3。其中,HINToAsp 的参数 ξ 的取值为 4.4 节中的最优取值,在大众点评数据集中 ξ 设置为 0.9, TripAdvisor 数据集中 ξ 设置为 0.85。

表 3 实验结果
Tab. 3 Experimental result

数据集	实验规模/%	RMSE					PCC				
		QPLSA	GRAOS	SATM	HINToAsp's	HINToAsp	QPLSA	GRAOS	SATM	HINToAsp's	HINToAsp
Dianping	25	0.5776	0.4825	0.5682	0.5723	0.4792	0.5634	0.1482	0.3589	0.5743	0.5675
	50	0.5723	0.4766	0.5646	0.5602	0.4736	0.5673	0.1502	0.3626	0.5755	0.5723
	75	0.5633	0.4702	0.5622	0.5563	0.4689	0.5682	0.1587	0.3798	0.5790	0.5804
	100	0.5545	0.4684	0.5597	0.5493	0.4625	0.5723	0.1632	0.3881	0.5792	0.5874
TripAdvisor	25	0.5732	0.5534	0.5547	0.5832	0.5413	0.4863	0.2056	0.4658	0.4687	0.5021
	50	0.5584	0.5423	0.5325	0.5729	0.5220	0.5133	0.2245	0.4765	0.4854	0.5123
	75	0.5327	0.5377	0.5210	0.5598	0.5187	0.5274	0.2476	0.4807	0.5143	0.5289
	100	0.5240	0.5328	0.5052	0.5377	0.4998	0.5372	0.2548	0.5032	0.5376	0.5418

由表 3 比较后发现,相对于对比的方法,在两个数据集合的各规模数据集上,HINToAsp 模型的 RMSE 的数值均最小,预测精度最高。而没有引入 HIN 的 HINToAsp's 效果和 QPLSA、SATM 模型的效果持平,比 GRAOS 算法效果差。据此,验证了引入结构信息能够有效提高评分预测的准确性,以及 HINToAsp 模型的有效性。此外,在大部分情况下,HINToAsp 模型预测结果的 PCC 值更好,将方面评分预测问题扩展为方面推荐问题时,HINToAsp 算法可以取得更好的效果,推荐的结果更接近真实排名。

综合分析表 3,尽管 GRAOS 在预测精度上效果也比较好,但是其 PCC 值是几个模型中最差的,而本文算法效果在两个指标上的效果均最好。

4.4 参数实验

在 HINToAsp 模型中,给定参数 ξ 调节主题挖掘模型和主题传播模型的贡献度。 ξ 取值为 0 至 1 中的实数。当 $\xi=1$ 时,主题传播模型不生效,只使用主题挖掘模型部分。参数 ξ 的取值由参数实验决定。实验结果见图 5。

由图 5(a)可知,大众点评数据集上, $\xi=0.9$ 时,取得最好的效果, $\xi=1$ 时,模型效果不是最好,因此可见,不能盲目使用结构信息。由图 5(b)可知, TripAdvisor 数据集上 $\xi=0.85$ 时取得最好预测效果。

5 结语

本文提出了一种基于异质信息网络和主题模型的方面分预测算法 HINToAsp。从内容信息和结构信息角度分别构建了基于 PLSA 的主题挖掘模型以及基于 HIN 的主题传播模

型;充分考虑了评论、评分等文本信息以及用户和商品之间构成的链接信息。本文通过和其他算法如 QPLSA、SATM 的对比,验证了 HINToAsp 算法的有效性。参数实验表明,恰当引入结构信息可以更加高效地进行评分预测和在评分预测基础上的推荐任务。

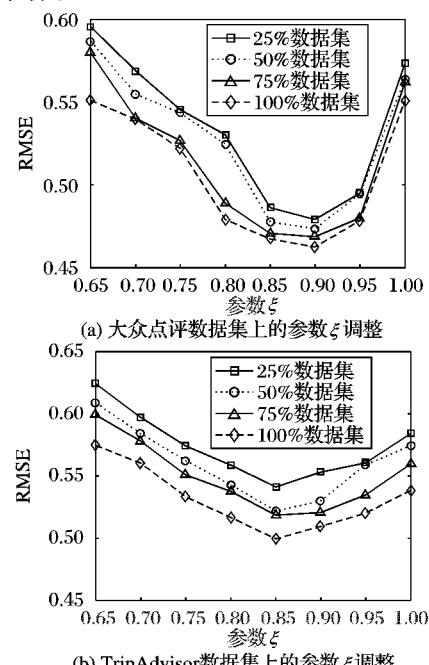


图 5 参数实验结果

Fig. 5 Parameter experimental results

参考文献 (References)

- [1] MOGHADDAM S, ESTER M. On the design of LDA models for as-



- pect-based opinion mining[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 803 – 812.
- [2] 林晓勇, 代苓苓, 史晟辉, 等. 基于主题模型的矩阵分解推荐算法[J]. 计算机应用, 2015, 35(S2): 122 – 124. (LIN X Y, DAI L L, SHI S H, et al. Matrix factorization recommendation based on topic model [J]. Journal of Computer Applications, 2015, 35(S2): 122 – 124.)
- [3] 王春龙, 张敬旭. 基于 LDA 的改进 K-means 算法在文本聚类中的应用[J]. 计算机应用, 2014, 34(1): 249 – 254. (WANG C L, ZHANG J X. Improved K-means algorithm based on latent Dirichlet allocation for text clustering [J]. Journal of Computer Applications, 2014, 34(1): 249 – 254.)
- [4] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 50 – 57.
- [5] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993 – 1022.
- [6] LU Y, ZHAI C X, SUNDARESAN N. Rated aspect summarization of short comments[C]// Proceedings of the 18th International Conference on World Wide Web. New York: ACM, 2009: 131 – 140.
- [7] SUN Y, HAN J, ZHAO P, et al. RankClus: integrating clustering with ranking for heterogeneous information network analysis[C]// Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 565 – 576.
- [8] SHI C, LI Y, ZHANG J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17 – 37.
- [9] ZHENG X, LIN Z, WANG X, et al. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification[J]. Knowledge-Based Systems, 2014, 61(2): 29 – 47.
- [10] WANG H, LU Y, ZHAI C X. Latent aspect rating analysis without aspect keyword supervision[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011: 618 – 626.
- [11] WANG H, ESTER M. A sentiment-aligned topic model for product aspect rating prediction[EB/OL].[2016-11-20]. <http://www.anthology.aclweb.org/D/D14/D14-1126.pdf>.
- [12] LI Y, SHI C, ZHAO H, et al. Aspect mining with rating bias [C]// Proceedings of the 2016 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer International Publishing, 2016: 458 – 474.
- [13] SHI C, ZHOU C, KONG X, et al. HeteRecom: a semantic-based recommendation system in heterogeneous networks[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1552 – 1555.
- [14] YU X, REN X, SUN Y, et al. Personalized entity recommendation: a heterogeneous information network approach[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM, 2014: 283 – 292.
- [15] SUN Y, HAN J. Mining heterogeneous information networks: a structural analysis approach [J]. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20 – 28.
- [16] 张邦佐, 桂欣, 何涛, 等. 一种融合异构信息网络和评分矩阵的推荐新算法[J]. 计算机研究与发展, 2014, 51(S2): 69 – 75. (ZHANG B Z, GUI X, HE T, et al. A novel recommender algorithm on fusion heterogeneous information network and rating matrix [J]. Journal of Computer Research and Development, 2014, 51 (S2): 69 – 75.)
- [17] LUO W, ZHUANG F, ZHAO W, et al. QPLSA: Utilizing quadruples for aspect identification and rating[J]. Information Processing and Management, 2015, 51(1): 25 – 41.
- [18] LUO W, ZHUANG F, CHENG X, et al. Ratable aspects over sentiments: predicting ratings for unrated reviews[C]// Proceedings of the 2014 IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2014: 380 – 389.

This work is partially supported by the National Natural Science Foundation of China (61375058), the National Basic Research Program (973 Program) of China (2013cb329606), the Co-construction Project of Beijing Municipal Commission of Education.

JI Yugang, born in 1993, Ph. D. candidate. His research interests include data mining, machine learning.

LI Yitong, born in 1992, M. S. Her research interests include data mining, machine learning.

SHI Chuan, born in 1978. Ph. D., professor. His research interests include data mining, machine learning, evolutionary computing.

ZHANG Xinming, born in 1963, M. S., professor. His research interests include intelligent optimization algorithm, digital image processing, pattern recognition.

KANG Qiang, born in 1989, M. S. candidate. His research interests include intelligent optimization algorithm, digital image processing.

WANG Xia, born in 1993, M. S. candidate. Her research interests include intelligent optimization algorithm, digital image processing.

CHENG Jinfeng, born in 1990, M. S. candidate. Her research interests include digital image processing.

(上接第 3200 页)

- [11] 李俊, 汪冲, 李波, 等. 基于扰动的精英反向学习粒子群优化算法[J]. 计算机应用研究, 2016, 33(9): 2584 – 2591. (LI J, WANG C, LI B, et al. Elite opposition-based particle swarm optimization based on disturbances [J]. Application Research of Computers, 2016, 33(9): 2584 – 2591.)
- [12] CHENG R, JIN Y C. A social learning particle swarm optimization algorithm for scalable optimization [J]. Information Sciences, 2015, 291(6): 43 – 60.

This work is partially supported by Key Scientific and Technologies Project of Henan Province (132102110209), the Research Program of Basic and Advanced Technology of Henan Province (142300410295).