计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：融合口碑和地理位置的竞争关系量化模型
作者：李艾鲜，乔少杰，韩楠，元昌安，黄萍，彭京，周凯
网络首发日期：2019-07-11
引用格式：李艾鲜，乔少杰，韩楠，元昌安，黄萍，彭京，周凯．融合口碑和地理位置的竞争关系量化模型．计算机科学与探索．
http://kns.cnki.net/kcms/detail/11.5602.TP.20190709.1329.017.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.1905094

融合口碑和地理位置的竞争关系量化模型*

李艾鲜[1]，乔少杰[2,3]，韩　楠[4+]，元昌安[5]，黄　萍[4]，彭　京[6]，周　凯[6]

1. 成都信息工程大学 网络空间安全学院，成都 610225
2. 成都信息工程大学 软件工程学院，成都 610225
3. 成都信息工程大学 软件自动生成与智能服务四川省重点实验室，成都 610225
4. 成都信息工程大学 管理学院，成都 610103
5. 南宁师范大学，南宁 530001
6. 四川省公安厅，成都 610014

+ 通讯作者 E-mail: hannan@cuit.edu.cn

摘　要：在同类服务或产品中识别和量化竞争是当前竞争关系挖掘领域关注的重要问题。本文提出科学合理的竞争关系评价指标，构建实体竞争关系综合评价指标体系，使用 LDA(Latent Dirichlet Allocation)模型对消费者口碑评论进行降维和主题提取，构建口碑相似度函数，对实体用户口碑相似度进行量化表示。根据实体地理位置属性，计算实体空间距离，构建实体相邻关系并以具有相邻关系实体的距离作为聚类中心，使用 KNN 算法对其进行聚类。综合上述技术提出 LTM(Location & Topical Model)模型，融合了用户评论、实体地理位置属性，量化实体间竞争关系。大量真实移动社交网络数据上实验结果表明所提方法的在量化指标制定、实用性和时间性能上具有较大优势。

关键词：竞争关系；关系量化；口碑；地理位置

文献标志码：A　**中图分类号：**TP311

李艾鲜，乔少杰，韩楠，等. 融合口碑和地理位置的竞争关系量化模型[J].计算机科学与探索
LI A X, QIAO S J, HAN N, et al. A Competitive Relationship Quantitative Model by Integrating Word of Mouth and Geographic Location[J]. Journal of Frontiers of Computer Science and Technology

A Competitive Relationship Quantitative Model by Integrating Word of Mouth and

* The National Natural Science Foundation of China under Grant Nos. 61772091, 61802035, 71701026 (国家自然科学基金); the Sichuan Science and Technology Program under Grant Nos. 2018JY0448, 2019YFG0106, 2019YFS0067, 2018GZ0114, 2018GZ0082, 2018GZ0307 (四川省科技计划项目); the Sichuan Major Science and Technology Special Program under Grant Nos. 2017GZDZX0002, 2018GZDZX0049 (四川省重大科技专项项目); the Innovative Research Team Construction Plan in Universities of Sichuan Province under Grant No. 18TD0027 (四川高校科研创新团队建设计划); the Natural Science Foundation of Guangxi under Grant No. 2018GXNSFDA138005 (广西自然科学基金项目); the Sichuan Science and Technology Innovation Seedling Project under Grant Nos. 2019016, 2019033 (四川省科技创新苗子工程项目); the Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology under Grant No. J201701 (成都信息工程大学中青年学术带头人科研基金); the Guangdong Key Laboratory Project under Grant No. 2017B030314073 (广东省重点实验室项目).

Geographic Location*

LI Aixian[1], QIAO Shaojie[2,3], HAN Nan[4+], YUAN Changan[5], HUANG Ping[4], PENG Jing[6], ZHOU Kai[6]

1. School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China
2. School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China
3. Software Automatic Generation and Intelligent Service Key Laboratory of Sichuan Province, Chengdu University of Information Technology, Chengdu 610225, China
4. School of Management, Chengdu University of Information Technology, Chengdu 610103, China
5. Nanning Normal University, Nanning 530001, China
6. Sichuan Provincial Department of Public Security, Chengdu 610014, China

Abstract: It is an important problem of identifying and quantifying the competition in similar services or products in the research area of competitive relationship mining. A scientific and reasonable evaluation metric of competitive relationship is proposed, and a comprehensive evaluation system of entity competitive relationship is constructed. Dimension reduction and theme extraction on users' reviews is achieved by using the LDA (Latent Dirichlet Allocation) model, the similarity function of comments is constructed, and the similarity degree of entity users' comments is quantified. Based on the geographic location information of entities, the spatial distance of entities is calculated, the adjacent relation of entities is constructed, and the distance of entities with adjacent relationship is regarded as the cluster center. The LTM (Location & Topical Model) model is proposed by integrating user's reviews, entity's geographical attributes, and quantifying the competitive relationship between entities. Experiments were conducted on a large number of real social network data of Yelp, and the results show that the proposed method has great advantages in quantitative metric formulation, practicability and time performance.

Key words: competitive relationship; relationship quantification; word-of-mouth; geographical location

1 引言

识别竞争对手、量化竞争关系是帮助企业、商家保持核心竞争力的重要方法[1]。现有研究通过专利挖掘[2, 3]、用户评论[4, 5]挖掘等方法识别竞争对手，鲜有竞争关系量化研究。翟东伟[6]构建主题-机构模型对专利机构的主题和竞争关系进行分析。Rodriguez 等人[7]提出了一种基于图形核的度量方法识别竞争对手。陈元[8]从 Web 用户评论中构建企业竞争情报挖掘模型获取企业产品竞争情报。聂卉等人[9]通过 Word2Vec 结合依存语法分析在线评论进行领域特征词典构建和用户观点抽取。上述工作仅实现了竞争情报的挖掘和竞争对手的识别，不能反映出实体间竞争关系的强弱。杨洋等人[10]提出主题因子图模型来量化推断企业间的竞争关系，但采用半监督学习方法，其实用性有限。上述研究均未考虑地理位置对竞争关系的影响，显然存在局限。

研究动机：口碑传播已被证明对消费者的购买决策起着重要的作用[11]。通过融合消费者口碑与地理位置信息进一步改进现有竞争关系挖掘方法，提升模型的实用性、客观性和准确性。电商企业可以将本文提出的新方法应用于评论特征抽取、评论内容中的企业竞争对手识别，将竞争关系发现与量化输出相结合，克服传统的竞争关系挖掘方法不考虑地理位置信息影响的缺点。融合消费者口碑和实体空间位置两大因素，科学地量化实体间竞争关系。

2 理论基础

本文中定义的实体包括但不限于企业、商店、餐厅等。首先给出竞争关系网络的定义，如下。

定义 1 竞争关系网络. 网络 $G = (V, E, S, L)$，V 是实体的集合，$E \subseteq V \times V$ 表示实体间的关系，S 表示该实体所有消费者的评论，L 代表实体的地理位置。

定义 2 实体主题模型. 实体的全部消费者评论集合 θ_d 的主题模型是单词$\{P(w|\theta_d)\}$的多项分布。一个餐厅 e_i 的所有消费者评论是从餐厅的主题模型 θ_d 中抽样形成的。

定义 3 困惑度[12]. 用来度量一个概率分布或概率模型预测结果的好坏程度，定义如下所示：

$$perplexity=e^{-\sum_w \log_2(p(w))/N} \quad (1)$$

其中，$p(w)$表示 LDA 模型中任意一个词 w 的概率，定义为：

$$p(w) = \sum_z p(z|d)p(w|z) \quad (2)$$

式 2 中，w 代表词，z 代表主题，d 代表文档，N 表示测试集中出现的所有词的数量(不排重)。$p(z|d)$表示从文档 d 抽取主题 z 的概率值，$p(w|z)$表示从主题 z 中抽取词 w 的概率值。因为 LDA 是词袋模型，困惑度是语料库的极大似然估计，即所有词的概率乘积，因此对于未知分布的数据集，其困惑度的值越小，说明主题模型越好，记录该条件下 LDA 主题模型取得的主题数量为 K(K 为最优值)。

定义 4 空间相邻关系. 当两个实体在地理空间中的最短路径小于或等于给定阈值 ξ 时，称两个实体空间相邻，用 *neighbor* 表示，定义如下：

$$neighbor(e_i, e_j) = distance(e_i, e_j) \le \xi \quad (3)$$

当空间中两个实体满足公式 3 时，说明空间中的实体对象 e_i 和 e_j 相邻。

3 竞争关系量化

本文基于消费者口碑(用户评论)和地理位置信息设计了 LTM(Location & Topical Model)模型，量化实体间竞争关系，辅助实体进行商业决策。

3.1 消费者口碑主题提取

消费者口碑是由消费者评论文本构成的文档数据，LDA 模型将主题视为词汇的概率分布，文档是主题的随机混合[13]。本文通过 LDA 主题模型提取实体消费者评论的主题与主题词。根据主题模型提出的主题和主题词分布，综合咨询专家意见和评价，建立“主题-特征”规则。依次对所有口碑评论进行规则匹配，统计规则匹配频率计算口碑相似度。

本文把实体 i 记为 e_i，其对应的所有消费者口碑评论视为一篇文档。假设有 n 个实体，那么对应 n 篇文档。假设有 K 个主题，则实体 i 的文本中的第 j 个词汇 w_{ij} 可以表示为：

$$p(w_{ij}) = \sum_z p(w_{ij} | z_i = k)p(z_i = k | d) \quad (4)$$

上式中 d 为 n 篇文档的集合，z_i 是潜在变量代表第 j 个词汇标签 w_{ij} 取自该主题，$p(w_{ij}|z_i)$是词汇 w_{ij} 属于主题 z_i 的概率，$p(z_i|d)$表示给定主题 z_i 属于当前文本的概率。

主题提取先统计 d 中出现过的词汇(不计重)W，制作词汇表，现假设 K 个主题形成 D 个文本以 W 个唯一性词汇表示，记=$p(w_{ij}|z_i=k)$为主题 z_i 下 W 个词汇的多项分布，其中 w_{ij} 是 W 个唯一性词汇表中的词汇。记=$p(z_i|d)$为文档 d 在 K 个主题上的多项分布。于是，文档 d 中词汇 w 的概率可表示为：

$$p(w|d) = \sum_{z_i}^{k} \vec{\varphi}_{z_i} \cdot \vec{\theta}_n \quad (5)$$

LDA 模型在上作 $Dirichlet(\vec{\alpha})$的先验概率假设，在上同样作 $Dirichlet(\vec{\beta})$的先验假设，得到 LDA 模型各层参数之间依赖关系的数学表述[14]如下：

$$z_i|\vec{\theta}_n \sim \text{Discrete}(\vec{\theta}_n),\ \vec{\theta}_n \sim Dirichlet(\vec{\alpha})$$

$$w_{ij}|z_i, \vec{\varphi}_k \sim \text{Discrete}(\vec{\varphi}_k),\ \vec{\varphi}_k \sim Dirichlet(\vec{\beta}).$$

LDA 主题提取模型需要给定数据集和主题的数量 K，根据定义 3 采用困惑度来确定 K 的取值。

3.2 消费者口碑相似度量化

在消费者口碑中，竞争关系越大的实体，其消费者的评论相似度越高。某商店消费者评论出现频率最高的词汇是“好喝”、“干净”、“服务”，其中“好喝”是针对奶茶口味，“干净”是针对设备，“服务”是针对店铺环境的。相似评论说明：在 A 商店消费

的消费者，有很大可能会在与 A 相似度高的 B 商店消费。因此需要对消费者口碑进行相似度量化。

根据主题模型建立“主题-特征”规则。依次对 n 篇文档利用公式 6 进行规则匹配。

$$score=\begin{cases}1 & (\text{匹配成功})\\ 0 & (\text{匹配失败})\end{cases} \tag{6}$$

“主题-特征”在本文档中出现则为匹配成功，否则为失败。以某一餐厅的评论为例，存在规则“food-nice”，则在该餐厅的所有用户口碑评论中搜索“food-nice”是否同时存在，若存在则匹配成功，$score$=1；反之失败，$score$=0。为了得每个实体的规则匹配分数，设计打分函数 $S(e_i)$：

$$Se_i=\sum_{r=1}^{R}\sum_{d=1}^{D}score_d \tag{7}$$

式 7 为统计匹配成功的频率，式中 e_i 代表第 i 个实体用户评论数据，作为函数的输入；r 代表规则数量；m 代表规则数量；n_i 表示 e_i 中词的数量；代表第 i 个实体匹配规则 r 后得到的分数。匹配完 m 个规则后，实体 i 获得一个分数 $score$。

$$sim_{ij}=e^{-|Se_i-Se_j|} \tag{8}$$

式 8 用于计算用户评论相似度。sim_{ij} 表示实体 i 与 j 的相似度。sim 值最小，说明实体相似度越大。

Table 1　Parameters and description of algorithm 1

表 1　算法 1 参数及说明

参数	说明
n	文档(实体)数量
D	文档集合
W	词汇集合
z	主题
K	主题集合
w	单词
e	实体
$\vec{\alpha}$	*Dirichlet* 参数
$\vec{\beta}$	*Dirichlet* 参数
$\vec{\theta_n}$	n 篇文档的主题分布
$\vec{\varphi_k}$	主题 k 的词分布

算法 1 消费者口碑量化算法

输入：$D;\vec{\alpha};\vec{\beta};K$.

输出：sim_{ij}.

1. **for each** i **in** D
2. ~ $Dirichlet(\vec{\alpha})$;
   //从 $Dirichlet(\vec{\alpha})$分布中抽样得主题分布
3. **end for**
4. **for each** k **in** K
5. $\vec{\theta_n}$ ~ $Dirichlet(\vec{\alpha})$;
   //从 $Dirichlet(\vec{\beta})$主题分布中抽样得到
6. **for each** j **in** W
7. $z_i|\vec{\theta_n}$ ~ $Discrete(\vec{\theta_n})$;
   //从多项式分布中抽样得到 $p(z_i|\vec{\theta_n})$
8. $w_{ij}|z_i,\vec{\varphi_k}$ ~ $Discrete(\vec{\varphi_k})$;
   //从多项式分布中抽样得到 $p(w_{ij}|z_i,\vec{\varphi_k})$
9. **end for**
10. **end for**
11. **for each** i **in** n
12. **for each** j **in** n
13. $sim_{ij}=e^{-|Se_i-Se_j|}$;
14. **end for**
15. **end for**

算法 1 的基本思想为：LDA 主题提取过程(第 1 行~第 10 行)，从参数为 $\vec{\alpha}$ 的 *Dirichlet* 分布中抽样生成第 i 个文档 n_i 的主题分布 $\vec{\theta_n}$；从参数为 $\vec{\beta}$ 的 *Dirichlet* 分布中抽样生成第 k 个主题的词分布 $\vec{\varphi_k}$；对于每一个词 w_{ij} 及其所属主题 z_i，首先从多项式分布 $\vec{\theta_n}$ 中抽样得到 $z_i=p(z_i|\vec{\theta_n})$，然后从多项式分布 $\vec{\varphi_k}$ 中抽样得到 $w_{ij}=p(w_{ij}|z_i,\vec{\varphi_k})$；求口碑相似度(第 11 行~第 15 行)。算法中的采样方法为 Gibbs 采样[15]。参数说明如表 1 所示。

时间复杂性分析：算法 1 时间复杂度为 O(K*N)，其中 K 表示主题数量，N 表示文档的总数。

3.3　地理位置相似度量化

本节设计了符合地理位置属性在实际生活中对竞争关系影响特点的相似度量化函数。dis 是距离矩阵，dis_{ij} 表示餐厅 i 与餐厅 j 之间的距离。算法的核心是将具有相似距离关系的餐厅聚集到一起，并赋予它们相同的影响因子 α，最终由实体距离影响力量化函数 $M(dis_{ij})$输出实体距离影响力量化结果。

根据定义 4，以存在相邻关系的实体 i、j 的相邻关系 $neighbor(e_i, e_j)$作为聚类的初始值，使用 KNN(K-Nearest Neighbor)算法对实体的经度纬度进行聚类得到 n 个簇，记为 C，$C=\{C_1, C_2, C_3, \ldots, C_n\}$。实际生活中，距离的远近将影响实竞争关系的强弱。把地理位置具有相似的点聚集到一起，同一个簇内，在地理位置属性上存在相似关系。不同的簇则相似性较弱。在互联网中，相距较远的实体也可能存在竞争关系。以美团为例，理发店 A 和理发店 B 相距五公里，但其主营业务一样，任然存在竞争关系。单纯的考虑距离来评价竞争关系会夸大距离对结果的影响，这显然是不合理的，因此本文引入地理位置属性影响因子 $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n\}$，定义如下：

$$a_i = \frac{C_i}{\Sigma_{i=1}^{n} C_i} \tag{9}$$

C_i 为簇 i 内点的数量，n 表示簇的数量。α 的作用包括：1) 调整距离对竞争关系的影响，强化簇内竞争关系，弱化簇间竞争关系；2) 给相似点赋相同的权重值，简化参数。由实体距离影响力量化函数：

$$M(dis_{ij}) = \alpha * dis_{ij} \tag{10}$$

输出实体距离影响力量化结果，式 10 中 dis_{ij} 表示一个二维矩阵，矩阵的行代表实体 e_i，矩阵的列代表实体 e_j，矩阵第 i 行第 j 列存放 e_i 到 e_j 的距离。$M(dis_{ij})$值越小，说明竞争关系越强。

3.4 LTM(Location & Topic Model)模型

在图 $\boldsymbol{G}=(V, E, S, L)$中，矩阵 E 中的值表示竞争关系的强弱。本文提出竞争关系量化函数 φ_{ij}，融合 2.1 及 2.2 节消费者口碑量化结果 sim_{ij}、实体地理位置属性影响力量化结果 $M(dis_{ij})$，其式为:

$$\varphi_{ij} = \frac{1}{sim_{ij} + M(dis_{ij})} \tag{11}$$

E_{ij} 表示实体 i 与实体 j 竞争关系归一化结果：

$$E_{ij} = \frac{\varphi_{ij} - \varphi_{min}}{\varphi_{max} - \varphi_{min}} \tag{12}$$

竞争关系量化算法参数说明如表 2 所示。

Table 2　Parameters and description of algorithm 2

表 2　算法 2 参数及说明

参数	说明
N	实体的数量
sim_{ij}	e_i 和 e_j 的相似度
$M(dis_{ij})$	e_i 和 e_j 的距离影响力
φ_{ij}	e_i 和 e_j 的竞争关系量化结果
E_{ij}	e_i 和 e_j 的竞争关系归一化结果

算法 2 竞争关系量化算法

输入: sim_{ij}; $M(dis_{ij})$。

输出: E_{ij}。

```
1.  for each i in N
2.      for each j in N
3.      φij= 1/(simij + M(disij));
4.      end for
5.  end for
6.  φmax = Find_max(φij);
7.  φmin = Find_min(φij);
8.  for each i in N
9.      for each j in N
10.     Eij = (φij −φmin)/(φmax -φmin);
11.     end for
12. end for
```

算法 2 工作原理：计算竞争关系量化值 φ_{ij}(第 1 行~第 5 行)；查找 φ_{ij} 中的最大值(第 6 行)，查找 φ_{ij} 中的最小值(第 7 行)；对竞争关系量化结果进行归一化处理(第 8 行~第 12 行)。

时间复杂性分析：通过分析算法 2，可知其时间复杂度为(N^2)，N 表示实体数量。

4　实验结果与分析

实验使用的数据为美国肯塔基州北部的城市 Louisville 地区 Yelp 网站上的餐厅数据，包含 2,375 个餐厅 ID 及其的地理位置属性和 66,156 条用户评论。实验硬件平台为：Intel(R) Core(TM) i5-4200M CPU 2.50GHz，操作系统平台为 Windows 10。

4.1　主题提取与相似度计算

在主题提取阶段，通过多次迭代得到困惑度变化曲线，并确定最佳主题数。实验中发现输入相同主题数，困惑度会有细微的波动。因此同一主题数采用多次实验取均值得到一条稳定的困惑度曲线。

如图 1 所示，当主题数量为 60 时，困惑度曲线

稳定收敛，说明该条件下模型对于实验数据集中的有效信息拟合较好，因此最佳的主题数取值为 60。

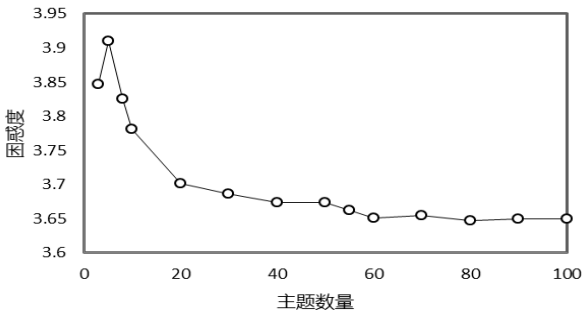

Fig.1　Curve of perplexity
图 1　困惑度曲线

根据主题提取结果，经过咨询领域专家，合并相似主题后，得到如表 3 所示的规则。

Table 3　Rules table of "Topic-Feature"
表 3　"主题-特征"规则表

主题	主题概率值	主题词(特征)	主题概率值
washed	0.000021823	clean	0.0040289415
equipment	0.040788829	contamination	0.0061728396
insect	0.130971804	food	0.0061728405
item	0.0096113039	temperature	0.1742978245
drinking	0.000021823	good	0.0288789775
nonfood	0.000021823	unclean	0.0322682150

Table 4　Rating score of rules matching
表 4　规则匹配评分表

餐厅名	ID	编号	分数
ANGIOS	29790	1	15
SAMARITAN	31334	2	8
CHEDDAR BOX	32112	3	13
CAFE AND PIZZERIA	32114	4	10
ARBYS	32140	5	5

通过 2.2 节的方法对数据集中的 2,375 个餐厅进行打分，本文以其中 5 家餐厅为例，结果如表 4 所示。根据式 8 计算餐厅之间的相似度，矩阵的行数表示 i 实体，列数表示 j 实体，sim_{ij} 表示餐厅 i 和餐厅 j 的消费者口碑相似度。

4.2　竞争关系量化

以表 4 所述餐厅为例根据定义 4 计算餐厅 i 与餐厅 j 之间的距离，实验将阈值 θ 设置为 1000 米，则数据中具有 *neighbor* 关系的点有 20 个。实验中采用 KNN 聚类算法，使用欧氏距离作为度量函数，把地理位置属性相似的餐厅聚为一类，重复 20 次，选聚类结果和 *neighbor* 关系点重合度最高的结果作为实验的聚类结果。根据聚类结果，由 2.3 节式 9 计算得到 α 值，其值是簇内的餐厅距离计算的权重，实验中簇与簇之间的 α 取 0.02。

根据式 10 计算餐厅竞争关系地理位置属性影响力量化结果 $M(dis_{ij})$。根据 3.4 节所提方法，得到最终的餐厅间竞争关系量化结果。可视化数据集中前 5 个餐厅之间的竞争关系，如图 2 所示。

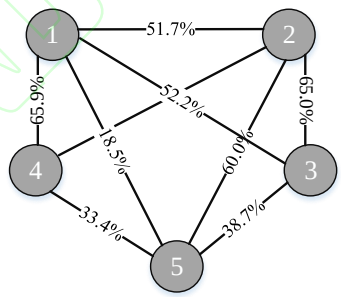

Fig.2　Visualized results of competitive relationship
图 2　竞争关系量化结果

使用仅考虑口碑对竞争关系影响的 TM(Topical model)模型进行对比实验，其结果如图 3 所示。

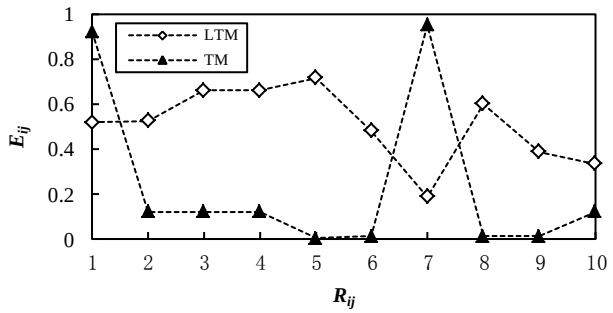

Fig.3　Comparison results of TM and LTM models
图 3　TM 与 LTM 模型对比实验结果

图 3 横轴 R_{ij} 代表餐厅 i 与餐厅 j 进行比较，纵轴 E_{ij} 代表餐厅 i 与餐厅的竞争关系量化结果，由于篇幅限制上图仅给出实验的前 10 个量化结果。通过图 3 可以发现：(1) TM 模型曲线波动很大，说明

仅仅靠用户口碑评论量化竞争关系容易会出现极端情况；(2) 以第 5 个点和第 6 个点为例，餐厅之间的竞争关系几乎为 0，这显然是不符合日常规律。因此仅仅靠口碑量化竞争关系是不准确的，因为同类餐厅的用户评论用词的重合度容易出现极端情况，不能很好的描述餐厅实际的竞争关系。图 3 中 LTM 模型在考虑地理位置属性后，对竞争关系的刻画符合实际情况。以 Yelp 网站而言，不论餐厅在城市的那个角度，都不应该出现竞争关系为 0 的情况，因为消费者完全可以驱车前往，即使是相距很远的餐厅也应该存在竞争关系。综上，LTM 模型能较好地刻画餐厅之间的竞争关系。

5 结束语

本文考虑消费者口碑和实体地理位置属性，提出 LTM 模型，量化表达消费者口碑和地理位置属性对实体竞争关系的影响。未来的研究工作包括：(1) 进一步挖掘实体竞争关系影响因素，例如时间属性对竞争关系的影响；(2) 现有竞争关系量化算法存在大量重复计算，设计新的算法降低时间复杂度，提升时间效率。

References：

[1] Porter M E. Competitive strategy: techniques for analyzing industries and competitors[J]. Social Science Electronic Publishing, 1980, 111(2): 86-87.

[2] Ma Y, Zhu L. Study on Patent Citation and firm Technical Competition[J]. Science of Science & Management of S & T, 2014, 35(3):42-49.

[3] Bao S H, Li R, Yu Y, et al. Competitor mining with the web[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(10): 1297–1310.

[4] Bao Z Y, Zhang H Q. Research of competitive intelligence mining based on patent map and case study[J]. Journal of Intelligence, 2016, 30(9): 20-23.

[5] Yang Y, Tang J, Li J Z. Learning to infer competitive relationships in heterogeneous networks[J]. ACM Transactions on Knowledge Discovery from Data, 2017, 12(1): 1-23.

[6] Qu D W. Reserch and application of patent information extraction and topic mining[D]. Beijing: Beijing University of Technology, 2017.

[7] Rodriguez A , Kim B , Lee J M, et al. Graph kernel based measure for evaluating the influence of patents in a patent citation network[J]. Expert Systems with Applications, 2015, 42(3): 1479-1486.

[8] Chen Y, Zhao J. The empirical study of enterprise competitive intelligence mining based on web user product reviews[J]. Information Science, 2016, 34(4): 80-85.

[9] Nie H, Li T, He H, et al. Automatic acquisition of business competition intelligence based on online reviews[J]. Journal of Intelligence, 2018, 37(10): 171-177.

[10] Yang Y, Tang J, Keomany J, et al. Mining competitive relationships by learning across heterogeneous networks[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Hawaii, October 29, 2012. New York: ACM, 2012: 1432-1441.

[11] Mikalef P, Pateli A, Giannakos M. Why are users of social media inclined to word of mouth?[C]//Proceedings of the 12th Theoretical Computer Science and General Issues Conference on E-Business, E-Services, E-Society, Athens, April 25-26, 2013. Berlin, Heidelberg: Springer, 2013: 112-123.

[12] Hoffman M D, Blei D M, Bach F. Online learning for latent dirichlet allocation[C]//Proceedings of 24th Annual Conference on Neural Information Processing Systems, Vancouver, December 6-9, 2010. Canada: Curran Associates, 2010: 865-864.

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[14] Steyvers M. Probabilistic topic models[J]. IEEE Signal Processing Magazine, 2010, 27(6): 55-65.

[15] Smith, A. F M, Roberts G O, et al. Bayesian computation via the Gibbs sampler and relater Markov chain Monte Carlo methods[J]. Journal of the Royal Statistical Society, 1993, 55(1): 3-23.

附中文参考文献：

[4] 鲍志彦，张红芹. 基于专利地图的竞争情报挖掘及实证研究[J]. 情报杂志, 2016, 30(9): 20-23.

[6] 翟东伟. 面向专利的信息抽取与主题挖掘技术研究及应用[D]. 北京: 北京工业大学, 2017.
[8] 陈元. 基于 WEB 用户产品评论的企业竞争情报挖掘实证研究[J]. 情报科学, 2016, 34(4): 80-85.
[9] 聂卉, 李通, 何欢, et al. 基于在线评论的商业竞争情报自动获[J]. 情报杂志, 2018, 37(10): 171-177.

LI Aixian was born in 1994. She is an M.S. candidate at Chengdu University of Information Technology, and member of CCF. Her research interests include Urban Computing, Data Mining, etc.
李艾鲜(1994–), 女, 四川宜宾人, 现为成都信息工程大学硕士研究生, CCF 会员, 主要研究领域为城市计算, 数据挖掘。

QIAO Shaojie was born in 1981. He received the Ph.D. degree from Sichuan University in 2009. He is a professor at Chengdu University of Information Technology. His research interests include mobile databases, trajectory prediction, machine learning and online social networks.
乔少杰(1981–), 男, 山东招远人, 博士, 教授, 主要研究领域为移动数据库, 轨迹预测, 机器学习, 社交网络。

HAN Nan was born in 1984. She received the Ph.D. degree from Chengdu University of Traditional Chinese Medicine in 2012. She is an associate professor at Chengdu University of Information Technology. Her research interests include trajectory prediction, online social networks and TCM data mining.
韩楠(1984–), 女, 陕西宝鸡人, 博士, 副教授, 主要研究领域为轨迹预测, 社交网络, 中医数据挖掘。

YUAN Changan was born in 1964. He received the Ph.D. degree from Sichuan University in 2006. He is a professor at Nanning Normal University. His research interests include databases and artificial intelligence.
元昌安(1964–), 男, 安徽肥东人, 博士, 教授, 主要研究领域为数据库, 人工智能。

HUANG Ping was born in 1963. She received the Ph.D. degree from Sichuan University in 2010. She is a professor at Chengdu University of Information Technology. Her research interests include tourism big data.
黄萍(1963-), 女, 四川成都人, 博士, 研究员, 主要研究领域为旅游大数据。

PENG Jing was born in 1973. He received the Ph.D. degree from Sichuan University in 2006. He is a senior research fellow at Sichuan Provincial Department of Public Security. His research interests include data mining.
彭京(1973-), 男, 四川成都人, 博士, 高级研究员, 主要研究领域为数据挖掘。

ZHOU Kai was born in 1984. He received the Master degree from Sichuan University in 2015. He is an engineer at Sichuan Provincial Department of Public Security. His research interests include information security, artificial intelligence, etc.
周凯(1984-), 男, 四川西昌人, 博士, 工程师, 主要研究领域为信息安全, 人工智能。