doi:10.11959/j.issn.1000-436x.2016125

基于受限非负张量分解的用户社会影响力分析

魏晶晶[1,2]，陈畅[3,4]，廖祥文[3,4]，陈国龙[3,4]，程学旗[5]

（1. 福州大学物理与信息工程学院，福建 福州 350116；2. 福建江夏学院电子信息科学学院，福建 福州 350108；
3. 福州大学数学与计算机科学学院，福建 福州 350116；4. 福州大学福建省网络计算与智能信息处理重点实验室，福建 福州 350116；
5. 中国科学院计算技术研究所，北京 100086）

摘 要：针对传统社会影响力分析方法未能充分考虑观点和话题信息等问题，提出了一种基于受限非负张量分解的用户社会影响力分析方法。首先把社交媒介用户相互评论关系自然地表示成三阶张量，然后通过拉普拉斯话题约束矩阵控制张量分解过程，最后根据分解得到的潜在因子度量用户观点社会影响力。该方法的优点是能有效地从受限张量分解结果中检索出给定话题下用户的社会影响力，同时保持其社会影响力的极性分布。实验结果表明，该方法的性能优于 OOLAM 和 TwitterRank 等基准算法。

关键词：社会影响力；话题；观点；张量分析

中图分类号：TP391 **文献标识码**：A

User social influence analysis based on constrained nonnegative tensor factorization

WEI Jing-jing[1,2], CHEN Chang[3,4], LIAO Xiang-wen[3,4], CHEN Guo-long[3,4], CHENG Xue-qi[5]

(1. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China;
2. College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou 350108, China;
3. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China;
4. Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350116,China;
5. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China)

Abstract: Existing models for measuring user social influence fail to integrate both opinion and topic information. Therefore, a new constrained nonnegative tensor factorization method combining user's opinion and the topical relevance was proposed. The method represented user's comment relations as 3-order tensor, factorized the comments tensor constrained by Laplacian topical matrix, and then measures user influence according to the latent factors resulting from the tensor factorization. Thus, the new method not only was capable to effectively calculate the strength of user social influence on given topic, but also kept the polarity allocation of social influence. The experimental result shows that the performance of the proposed method is better than that of the baseline methods such as OOLAM , TwitterRank, etc.

Key words: social influence, topic, opinion, tensor analysis

1 引言

社会影响力是指一个人的思想、情感或行为被他人所影响的现象[1,2]，其作为一种影响网络结构和信息传播的重要因素，受到了许多研究者的关注。社会影响力分析往往通过分析人们的社会交互行为来研究人们的社会影响，并在多个研究领域中起到关键作用，如推荐系统[3]、社交网络信息传播[4,5]、

收稿日期：2015-05-22；**修回日期**：2016-01-30
通信作者：廖祥文，liaoxw@fzu.edu.cn
基金项目：国家自然科学基金资助项目（No.61300105）；教育部博士点联合基金资助项目（No.20123514l0010）；福建省科技重大专项基金资助项目（No.2013H6012）；福州市科技计划基金资助项目（No.2012-G-113, No.2013-PT-45）

Foundation Items: The National Natural Science Foundation of China (No.61300105), The Research Fund for Doctoral Program of Higher Education of China (No.20123514l0010), The Key Project of Science and Technology of Fujian (No.2013H6012), The Project of Science and Technology of Fuzhou (No.2012-G-113, No.2013-PT-45)

突发事件检测[6]和广告投放[7]等。

在线社交网络出现和兴起之前，针对社会影响力的研究工作主要集中在理论层面，包括二级传播理论、弱连带优势理论、强连带优势理论和结构洞理论等[8]。随着微博、Facebook 等社交媒体广泛使用，人们可以在社交媒介上随时随地发布信息，而不受时间和空间的限制。这些海量的用户自创造数据(user generated data)蕴含非常丰富的用户信息，如用户观点、用户间交互关系等，为社会影响力分析理论的验证与应用提供了理想的环境。从内容角度，社会影响力分析可分为 3 方面[9]：1) 社会影响力自身的识别，研究影响力和相关因素的联系；2) 社会影响力的度量，希望能够找到合适的度量社会影响力的方法；3) 社会影响力的动态传播，即刻画社会影响力的动态特性。社会影响力的度量方法主要有 4 个角度[9]：1）基于网络拓扑结构的度量，通过衡量网络图中节点与连接的重要性来体现社会影响力的大小；2）基于用户行为的度量，使用统计等方法分析用户在社交网络中留下的行为数据；3）基于用户交互信息的度量，主要包括基于交互信息内容的度量和基于话题的度量；4）基于时间因素、转移熵等其他度量。

从层次角度，社交影响力分析主要有以下 3 个层次。1）整体社交影响力分析，毛佳昕等[8]提出用户关注、微博转发这 2 种用户行为与时间维度有关，以及转发延迟的分布近似服从幂律分布 2 个假设，并通过假设检验验证，最后使用全局阅读期望的方法度量用户影响力。2）话题级社交影响力分析，Weng 等[10]提出了一种结合网络结构与话题信息来计算话题级社会影响力的方法，验证了话题相似的用户间更容易互相产生影响。据此，在 PageRank 基础上加入话题相似度的因素，提出了一种 TwitterRank 方法并取得了不错的效果。3）信息条目级社交影响力分析：Cui 等[1,2]提出了一种更细粒度的社交影响力度量思路，即信息条目级社会影响力度量。其使用受限非负矩阵分解的方法来预测用户在某一话题下的社会影响力大小，矩阵约束的部分考虑了用户朋友活跃度、用户与朋友关系强度以及话题信息，该方法的实验效果较好。

当前，细粒度的社会影响力分析更加引起了研究者的重视，用户观点已成为度量用户社会影响力不可忽视的因素。另一方面，用户社会影响力与话题密切相关。Cai 等[11]曾提出利用带有倾向性连接的网络度量用户的社会影响力，并提出了一种可并行化的 PageRank 改进方法来求解所提出的 OOLAM 模型，得到 2 个独立的用户正负面影响力评分，从而更加细致地刻画了社会影响力。然而，该方法不能很好地融入用户的话题信息，难以分析领域专家的社会影响力。Weng 等[10]提出的 TwitterRank 方法将话题信息融入到用户社会影响力分析中，能够有效地检索出给定话题下比较重要的用户，但是却不能反映出用户社会影响力的正负面倾向。导致这一局限性的根本原因在于基于图的方法主要是刻画二维数据，难以同时将不同的信息加入到分析过程中。张量[12]是一种特别适合表达多维数据、融合不同信息的数据表达方式，广泛应用于多模态特征融合相关研究。

因此，本文提出一种基于受限非负张量分解的用户观点社会影响力分析方法，度量特定话题下用户的社会影响力及其影响力的极性分布。该方法首先使用张量表示用户相互评论关系，然后通过 Laplacian 矩阵将话题信息融入到张量分解中，最后基于分解得到的潜在因子度量在特定话题下用户观点的社会影响力。通过实验表明，本文方法不仅在效果上比 OOLAM、TwitterRank 等方法有一定的提升，而且能够更加细致地刻画用户观点的社会影响力。

2　用户观点社会影响力估计模型

2.1　问题描述

在社交媒介上，若标记用户集为 $U=\{u_1, u_2,\cdots,u_n\}$，其中，每个用户包含其所发表的文档集 $Doc=\{d_1,d_2,\cdots,d_i\}$，用户链接集 $Lin=\{l_1,l_2,\cdots,l_j\}$ 等信息；每篇文档 d 关联着评论集 $C=\{c_1,c_2,\cdots,c_k\}$ 等信息。用户观点的社会影响力分析的任务可表示为给定话题 q，寻找在话题 q 下，用户观点 o 的社会影响力度量函数 $f_o(U,Doc,q,C)\to\hat{S}$，其中，f_o 表示一个映射关系，$\hat{S}\in R$ 是用户观点 o 的社会影响力估计值。其中，o 取值可为 1、0 或−1，分别代表正面、中性、负面观点。

用户观点 o 是社会影响力的真实集合，记为 $S=\{s_1,s_2,\mathrm{L}\ ,s_n\}$，用户观点 o 的影响力估计值集合记为 $\hat{S}=\{\hat{s}_1,\hat{s}_2,\mathrm{L}\ ,\hat{s}_n\}$。用户观点 o 社会影响力的度量可以看成是在话题 q 的条件下，找到一个映射 f_o 使 $\hat{S}$ 与 S 的偏差达到最小。

2.2 张量代数介绍

遵循 Kolda 和 Bader 的符号描述，简要介绍与本文工作相关的张量代数基本知识[12]。

定义 1 张量 $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ 的 PARAFAC 分解形式为 $\boldsymbol{X} \approx \sum_{r=1}^{R} \lambda_r \boldsymbol{A}_r^{(1)} \circ \boldsymbol{A}_r^{(2)} \circ \cdots \circ \boldsymbol{A}_r^{(N)}$，$\boldsymbol{\lambda} \in \mathbb{R}^{R}$，‘$\circ$’表示外积，$\boldsymbol{A}^{(n)} \in \mathbb{R}^{I_n \times R}$，对于 $n = 1, 2, \cdots, N$，$\boldsymbol{A}_r^{(n)}$ 表示 $\boldsymbol{A}^{(n)}$ 的第 r 列。

定义 2 张量 $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ 在 n 模式下展开的符号为 $X_{(n)} \in \mathbb{R}^{I_n \times (I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N)}$，其展开方法的运算规则为

$$(X_{(n)})_{i_n j} = X_{i_1 i_2 \cdots i_N}, j = 1 + \sum_{k=1, k \neq n}^{N} (i_k - 1) J_k, J_k = \prod_{m=1, m \neq n}^{k-1} I_m \tag{1}$$

定义 3 矩阵 $\boldsymbol{A} \in \mathbb{R}^{I \times J}$ 与矩阵 $\boldsymbol{B} \in \mathbb{R}^{K \times L}$ 的 Kronecker 乘法运算规则为

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1J}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2J}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\boldsymbol{B} & a_{I2}\boldsymbol{B} & \cdots & a_{IJ}\boldsymbol{B} \end{bmatrix} \tag{2}$$

定义 4 矩阵 $\boldsymbol{A} \in \mathbb{R}^{I \times K}$ 与矩阵 $\boldsymbol{B} \in \mathbb{R}^{J \times K}$ 的 Katri-Rao 乘法规则为

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} a_1 \otimes b_1 & a_2 \otimes b_2 & \cdots & a_K \otimes b_K \end{bmatrix} \tag{3}$$

2.3 基于受限非负张量分解的用户观点社会影响力分析方法

在应用的驱动下，越来越多研究工作关注话题级或条目级等更加细致的用户社会影响力分析。本文所关注的问题是分析特定话题下用户观点的社会影响力和极性分布。通过观察，本文发现：1）与话题相关度高的用户往往越容易获得其他用户的评论，其收到的评论总量一般会高出与话题无关的用户；2）话题相关的用户所发布的文档往往采用分布类似的词来描述话题。基于用户话题相似性特征，本文提出了一种基于受限非负张量的方法。该方法首先利用张量自然地对用户之间的评论关系建模，然后通过加入用户话题相似矩阵控制张量分解过程，最后基于张量分解得到的潜在因子度量用户观点的社会影响力和观点极性分布。

2.3.1 基于用户评论关系的张量构建

用户与用户之间带有观点评论的三元关系，可以用一个三阶张量 $\boldsymbol{X} \in \{0,1\}^{N \times M \times E}$ 刻画用户间的评论行为。其中，张量的 1 模式表示被评论用户，2 模式表示发表评论的用户，3 模式表示评论的观点倾向性，倾向性分为正面、中性、负面 3 种情况。这里的模式对应张量的每一个维度。每个张量元素值为

$$X_{ijk} = \begin{cases} 1, u_j \text{对} u_i \text{进行了观点为} k \text{的评价} \\ 0, \text{其他} \end{cases} \tag{4}$$

需要说明的是，判定用户 u_j 对用户 u_i 的评价观点，即观点倾向性的极性，是通过基于情感词典[13]的判定方法获得的。若评价内容中正面情感词数大于负面情感词数，则记为一次正面观点的评价，若评价内容中正面情感词数等于负面情感词数，则记为一次中性观点的评价，否则记为一次负面观点的评价。

2.3.2 用户话题相似性计算

首先，给定话题 q，用户 u_i 及其发布过的文档集合 $D_{u_i} = \{d_1, d_2, \cdots, d_m\}$。计算 u_i 发布过的文档的 BM 25（best match 25）值。在给定 BM 25 阈值的情况下，则可以得到 $subD_{u_i} = \{d_1, d_2, \cdots, d_l\}$，即 BM 25 值大于阈值的文档子集。然后，运用 LDA（latent dirichlet allocation）模型计算所有用户 $subD$ 的话题词集 $T = \{t_1, t_2, \cdots, t_k\}$。

通过 LDA 模型得到话题词集 $T = \{t_1, t_2, \cdots, t_k\}$ 之后，可以计算 u_i 在话题词集下的特征向量 $\vec{\boldsymbol{u}}_i$。该特征向量采用词袋模型（bag of words model）。这种模型把文本（段落或者文档）看作是无序的词汇集合，忽略语法甚至是单词的顺序，是自然语言处理和信息检索领域的一种常用假设。因此以词袋向量作为用户的特征向量，可以使用类似 cosine 相似度的计算方法计算用户相似度，计算公式为

$$Sim(u_i, u_j) = \frac{\vec{\boldsymbol{u}}_i \cdot \vec{\boldsymbol{u}}_j}{\left\| \vec{\boldsymbol{u}}_i \right\|^2 \left\| \vec{\boldsymbol{u}}_j \right\|^2} \tag{5}$$

其中，运算 $\cdot$ 表示向量内积，运算 $\|\cdot\|$ 表示向量模长。由于领域专家将包含更多领域专业词，因而公式中分母是模长平方的乘积，这使领域专家与其他普通用户的相似度将更低，更好地与非领域专家区分开。然后给定阈值 θ，则可以得到用户话题相似矩阵 $\boldsymbol{H} \in \{0,1\}^{n \times n}$，计算方法为

$$\boldsymbol{H}_{ij} = \begin{cases} 1, Sim(u_i, u_j) \geqslant \theta \\ 0, Sim(u_i, u_j) < \theta \end{cases} \tag{6}$$

2.3.3 改进的受限非负张量分解方法

针对评论关系张量，根据用户话题相似性假

设，提出一种 CP(CANDECOMP/PARAFAC)分解算法 CP_ALS[14]的改进算法 HF-CP-ALS，并通过该算法分解得到刻画用户观点社会影响力的潜在因子矩阵。

对给定用户间评论关系的张量表示为 $\boldsymbol{X}\in\{0,1\}^{N\times M\times 3}$，对应的 PARAFAC 分解最优化目标函数为

$$\min_{\boldsymbol{\lambda},\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\boldsymbol{A}^{(3)}}\left\|\boldsymbol{X}-\sum_{r=1}^{R}\boldsymbol{\lambda}_r\boldsymbol{A}_r^{(1)}\circ\boldsymbol{A}_r^{(2)}\circ\boldsymbol{A}_r^{(3)}\right\|_{\mathrm{F}}^2 \tag{7}$$

其中，$\boldsymbol{\lambda}\in\mathbb{R}^R$ 表示权重向量，$\boldsymbol{A}^{(i)}$ 表示第 i 个模式的潜在因子矩阵，$\boldsymbol{A}_r^{(i)}$ 表示 $\boldsymbol{A}^{(i)}$ 的第 r 列，符号‘o’表示外积，$\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数。

为求解目标函数式(7)，先求解在 CP_ALS 算法中的 1 模式最优化目标函数为

$$\begin{aligned}&\min_{\hat{\boldsymbol{A}}^{(1)}}\left\|\boldsymbol{X}_{(1)}-\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}\right\|_{\mathrm{F}}^2\\&\text{s.t. }\hat{\boldsymbol{A}}=\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{A}^{(1)},\boldsymbol{Y}=\boldsymbol{A}^{(3)}\ \mathrm{e}\ \boldsymbol{A}^{(2)}\end{aligned} \tag{8}$$

其中，$\boldsymbol{\lambda}\in\mathbb{R}^R$ 表示权重向量，$\boldsymbol{A}^{(i)}$ 表示第 i 个模式的潜在因子矩阵，$\mathrm{diag}(\boldsymbol{\lambda})$ 表示以 $\boldsymbol{\lambda}$ 为对角元素的方阵，运算符“e”表示 Katri-Rao 乘法。

在 CP_ALS 算法 1 模式的最优化目标函数中加入用户话题相似性限制，从而获得限定话题下的用户观点社会影响力。在该约束下，话题相关而且影响力小的那些用户，其用户观点社会影响力将提升，对于那些话题无关而且影响力大的用户，其用户观点社会影响力将减小。此外，为了保证潜在因子的可解释性，引入 $\hat{\boldsymbol{A}}^{(1)}\geqslant 0$ 的约束，得到

$$\begin{aligned}&J=\min_{\boldsymbol{\lambda},\boldsymbol{A}^{(1)}}\left\|\boldsymbol{X}_{(1)}-\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\alpha\sum_{ij}\boldsymbol{H}_{ij}\left\|\hat{\boldsymbol{X}}_{(1)i*}-\hat{\boldsymbol{X}}_{(1)j*}\right\|_{\mathrm{F}}^2\\&\text{s.t. }\hat{\boldsymbol{A}}^{(1)}\geqslant 0,\ \hat{\boldsymbol{A}}^{(1)}=\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{A}^{(1)},\boldsymbol{Y}=\boldsymbol{A}^{(3)}\ \mathrm{e}\ \boldsymbol{A}^{(2)}\end{aligned} \tag{9}$$

其中，$\boldsymbol{\lambda}\in\mathbb{R}^R$ 表示权重向量，$\boldsymbol{A}^{(i)}$ 表示第 i 个模式的潜在因子矩阵，$\mathrm{diag}(\boldsymbol{\lambda})$ 表示以 $\boldsymbol{\lambda}$ 为对角元素的方阵，$\boldsymbol{H}$ 表示用户相似性矩阵，$\hat{\boldsymbol{X}}$ 表示对 $\boldsymbol{X}$ 的估计即 $\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}$，$\alpha$ 是调节因子，运算符“e”表示 Katri-Rao 乘法。

用户评论张量 $\boldsymbol{X}\in\{0,1\}^{N\times M\times 3}$，其展开示意如图 1 所示。$\boldsymbol{X}_{(1)}$ 是一个 $N\times 3\times M$ 的矩阵，如图 2 所示，根据用户话题相似性特征，加入用户话题相似性约束后，话题相似用户 u_i,u_j 在 $\boldsymbol{X}_{(1)}$ 中的值将相互接近，即向量 $\hat{\boldsymbol{X}}_{(1)i*}$ 与向量 $\hat{\boldsymbol{X}}_{(1)j*}$ 会相互接近，从而让话题相关而且影响力小的用户获得更高的影响力。同时降低话题无关而且影响力大的用户的影响力。虽然该方法可能使那些话题即相关影响力高的用户的影响力略微下降，但是不会影响其整体排序。

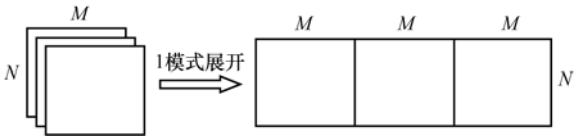


图 1　三阶张量 1 模式展开示意

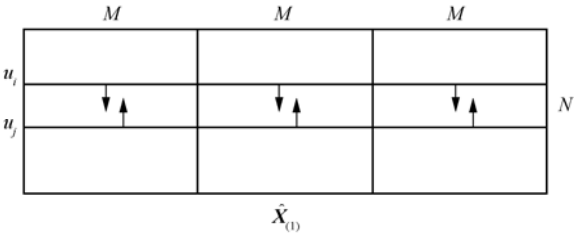


图 2　用户话题相似限制影响分解过程示意

直接求解式（9）所描述的优化问题时间复杂度过高，为简化运算引入拉普拉斯矩阵[15]$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{H}$，$\boldsymbol{D}\in\mathbb{R}^{N\times N}$。$\boldsymbol{D}$ 是一个对角矩阵，$\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{H}_{ij}$。由于 $\boldsymbol{L}$ 近似为对角占优矩阵，因此用 $\boldsymbol{D}$ 近似 $\boldsymbol{L}$，可以得到

$$\begin{aligned}\sum_{ij}\boldsymbol{H}_{ij}\left\|\hat{\boldsymbol{X}}_{(1)i*}-\hat{\boldsymbol{X}}_{(1)j*}\right\|_{\mathrm{F}}^2&=\sum_{k=1}^{c}\hat{\boldsymbol{X}}_{(1)*k}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{H})\hat{\boldsymbol{X}}_{*k}\\&\approx\mathrm{tr}(\boldsymbol{Y}\hat{\boldsymbol{A}}^{(1)\mathrm{T}}\boldsymbol{D}\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}})\end{aligned} \tag{10}$$

其中，$\mathrm{tr}(\cdot)$ 表示矩阵的迹。

引入拉普拉斯矩阵后，待优化的目标函数可以写成如下形式

$$\begin{aligned}&J=\min_{\boldsymbol{\lambda},\boldsymbol{A}^{(1)}}\left\|\boldsymbol{X}_{(1)}-\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\alpha\mathrm{tr}(\boldsymbol{Y}\hat{\boldsymbol{A}}^{(1)\mathrm{T}}\boldsymbol{D}\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}})\\&\text{s.t. }\hat{\boldsymbol{A}}^{(1)}\geqslant 0,\ \hat{\boldsymbol{A}}^{(1)}=\mathrm{diag}(\lambda)\boldsymbol{A}^{(1)}\\&\boldsymbol{Y}=\boldsymbol{A}^{(3)}\ \mathrm{e}\ \boldsymbol{A}^{(2)}\end{aligned} \tag{11}$$

张量分解中解决该类型的优化问题常用交替最小二乘法（ALS）求解目标函数 J，即更新其中一个因子矩阵时固定另外 2 个因子矩阵。α 表示限制项的重要程度，因此先计算 J 对 $\hat{\boldsymbol{A}}^{(1)}$ 的微分

$$\begin{aligned}&J(\hat{\boldsymbol{A}}^{(1)})=\mathrm{tr}((\boldsymbol{X}_{(1)}-\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X}_{(1)}-\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}))+\\&\alpha\mathrm{tr}(\boldsymbol{Y}\hat{\boldsymbol{A}}^{(1)\mathrm{T}}\boldsymbol{D}\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}})\\&\Rightarrow\frac{\partial}{\partial\hat{\boldsymbol{A}}^{(1)}}J=-2\boldsymbol{X}_{(1)}\boldsymbol{Y}+2\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}+2\alpha\boldsymbol{D}\hat{\boldsymbol{A}}^{(1)}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\end{aligned} \tag{12}$$

通过计算驻点来确定 $\hat{\boldsymbol{A}}^{(1)}$ 的值，令 $\frac{\partial}{\partial \hat{\boldsymbol{A}}^{(1)}} J = 0$，整理后就可以容易地计算得到 $\hat{\boldsymbol{A}}^{(1)}$ 的更新规则为

$$\hat{\boldsymbol{A}}^{(1)} = (\boldsymbol{I} + \alpha \boldsymbol{D})^{\dagger} \boldsymbol{X}_{(1)} \boldsymbol{Y} (\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y})^{\dagger} \tag{13}$$

其中，符号“†”表示矩阵的伪逆，$\boldsymbol{I}$ 是单位矩阵。对于剩下的另外 2 个潜在因子矩阵 $\hat{\boldsymbol{A}}^{(2)}$、$\hat{\boldsymbol{A}}^{(3)}$，其更新规则与算法 CP_ALS 的相同，更新规则如下。

$$\begin{cases} \hat{\boldsymbol{A}}^{(2)} = \boldsymbol{X}_{(2)} \left(\boldsymbol{A}^{(3)} \mathrm{e} \ \boldsymbol{A}^{(1)} \right) \left(\boldsymbol{A}^{(3)\mathrm{T}} \boldsymbol{A}^{(3)} \boldsymbol{A}^{(1)\mathrm{T}} \boldsymbol{A}^{(1)} \right)^{\dagger} \\ \hat{\boldsymbol{A}}^{(3)} = \boldsymbol{X}_{(3)} \left(\boldsymbol{A}^{(2)} \mathrm{e} \ \boldsymbol{A}^{(1)} \right) \left(\boldsymbol{A}^{(2)\mathrm{T}} \boldsymbol{A}^{(2)} \boldsymbol{A}^{(1)\mathrm{T}} \boldsymbol{A}^{(1)} \right)^{\dagger} \end{cases} \tag{14}$$

至此已经得到了 3 个潜在因子矩阵的更新规则，加入非负性约束后可以得到算法 HF-CP-ALS，其伪代码如图 3 所示。

Procedure HF-CP-ALS($\boldsymbol{X}$,$\boldsymbol{D}$,$\boldsymbol{R}$)
初始化 $\boldsymbol{A}^{(n)} = \in \mathrm{R}^{I_n \times R}, n = 1,2,3$
Repeat
$\hat{\boldsymbol{A}}^{(1)} = (\boldsymbol{I} + \alpha \boldsymbol{D})^{\dagger} \boldsymbol{X}_{(1)} \boldsymbol{Y} (\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y})^{\dagger}$
单位化 $\hat{\boldsymbol{A}}^{(1)}$ 的每一列，将 $\hat{\boldsymbol{A}}^{(1)}$ 中小于 0 的值置零，更新$\boldsymbol{\lambda}$
$\hat{\boldsymbol{A}}^{(2)} = \boldsymbol{X}_{(2)} \left(\boldsymbol{A}^{(3)} \mathrm{e} \ \boldsymbol{A}^{(1)} \right) \left(\boldsymbol{A}^{(3)\mathrm{T}} \boldsymbol{A}^{(3)} \boldsymbol{A}^{(1)\mathrm{T}} \boldsymbol{A}^{(1)} \right)^{\dagger}$
单位化 $\hat{\boldsymbol{A}}^{(2)}$ 的每一列，将 $\hat{\boldsymbol{A}}^{(2)}$ 中小于 0 的值置零，更新$\boldsymbol{\lambda}$
$\hat{\boldsymbol{A}}^{(3)} = \boldsymbol{X}_{(3)} \left(\boldsymbol{A}^{(2)} \mathrm{e} \ \boldsymbol{A}^{(1)} \right) \left(\boldsymbol{A}^{(2)\mathrm{T}} \boldsymbol{A}^{(2)} \boldsymbol{A}^{(1)\mathrm{T}} \boldsymbol{A}^{(1)} \right)^{\dagger}$
单位化 $\hat{\boldsymbol{A}}^{(3)}$ 的每一列，将 $\hat{\boldsymbol{A}}^{(3)}$ 中小于 0 的值置零，更新$\boldsymbol{\lambda}$
Until 收敛或达到最大迭代次数
return $\boldsymbol{\lambda}$，$\boldsymbol{A}^{(1)}$，$\boldsymbol{A}^{(2)}$，$\boldsymbol{A}^{(3)}$
end procedure

图 3 受限张量分解 HF-CP-ALS 算法

在算法 HF-CP-ALS 中，值得注意的是在每一次更新因子矩阵完毕后，需要对矩阵做一次列向量单位化。特别地，潜在因子矩阵具有非负性约束，因此，在更新完 $\boldsymbol{A}^{(1)}$、$\boldsymbol{A}^{(2)}$或 $\boldsymbol{A}^{(3)}$时还需将其中小于零的元素置为 0，从而保持潜在因子矩阵非负，即保证潜在因子矩阵的可解释性。最后同时更新向量 $\boldsymbol{\lambda}$。HF-CP-ALS 算法最终可以求得各个模式的潜在因子矩阵和向量 $\boldsymbol{\lambda}$。

2.3.4 *用户观点社会影响力度量*

用户观点的社会影响力往往由一系列潜在因子决定，可通过分析潜在特征矩阵计算得到[16,17]。通过算法 HF-CP-ALS 容易得到话题约束下的用户观点潜在因子：$\boldsymbol{A}_r^{(1)}$、$\boldsymbol{A}_r^{(2)}$ 和 $\boldsymbol{A}_r^{(3)}$。设 $|\boldsymbol{\lambda}|$ 表示向量 $\boldsymbol{\lambda}$ 的长度，那么分解结果可以看成 $|\boldsymbol{\lambda}|$ 个秩为一的张量之和，其计算式可以写成

$$\boldsymbol{X} \approx \sum_{r=1}^{|\boldsymbol{\lambda}|} \boldsymbol{\lambda}_r \boldsymbol{A}_r^{(1)} \circ \boldsymbol{A}_r^{(2)} \circ \boldsymbol{A}_r^{(3)} \tag{15}$$

设选定某一 $\boldsymbol{\lambda}_k$，那么对应的秩 1 张量可以表示为向量的外积 $\boldsymbol{X}_k = \boldsymbol{\lambda}_k \boldsymbol{A}_k^{(1)} \circ \boldsymbol{A}_k^{(2)} \circ \boldsymbol{A}_k^{(3)}$。其中，$\boldsymbol{\lambda}_k$ 表示 $\boldsymbol{X}_k$ 对 $\boldsymbol{X}$ 的重要程度。令 $\boldsymbol{Z}_k = \boldsymbol{A}_k^{(1)} \circ \boldsymbol{A}_k^{(2)}$，则任意用户 u_j 在 $\boldsymbol{\lambda}_k$ 下的综合影响力得分计算式为 $\boldsymbol{S}_j = \boldsymbol{\lambda}_k \sum_i (\boldsymbol{Z}_k)_{ji}$。在 2.3.1 节中已约定 3 模式的每个维度分别表示正面、中性、负面观点。因此，正面影响力的计算方法为 $\boldsymbol{S}_j^{+} = \boldsymbol{S}_j \left(\boldsymbol{A}_k^{(3)} \right)_1$，负面影响力得分为 $\boldsymbol{S}_j^{-} = \boldsymbol{S}_j \left(\boldsymbol{A}_k^{(3)} \right)_3$，所有用户收到的所有评论的极性强度分布就是 $\boldsymbol{A}_k^{(3)}$。$\boldsymbol{A}_k^{(3)}$ 是一个向量，$\left(\boldsymbol{A}_k^{(3)} \right)_i$ 表示其第 i 个元素。

在计算某一 $\boldsymbol{\lambda}_k$ 下的用户观点影响力后，就可以完成计算用户在某一话题 q 下的用户观点社会影响力。首先，对 $\boldsymbol{\lambda}$ 中的值降序排序得到 $\boldsymbol{\lambda}'$，对应潜在因子矩阵的列也应根据 $\boldsymbol{\lambda}'$ 调整位置得到 $\boldsymbol{A}^{(n')}$，然后确定前 l 大的 $\boldsymbol{\lambda}_k$，l 应满足 $\frac{\sum_{i=1}^{l} \boldsymbol{\lambda}_i'}{\sum_{i=1}^{R} \boldsymbol{\lambda}_i} \geqslant \varepsilon$。最后用户 u_j 的社会影响力得分 $\boldsymbol{S}_j$ 的计算方法为

$$\boldsymbol{S}_j = \sum_{k=1}^{l} \boldsymbol{\lambda}_k \sum_i \left(\boldsymbol{A}_k^{(1)'} \circ \boldsymbol{A}_k^{(2)'} \right)_{ji} \tag{16}$$

那么，用户 u_j 极性是 p 的社会影响力为

$$\boldsymbol{S}_j = \sum_{k=1}^{l} \boldsymbol{\lambda}_k' \left(\boldsymbol{A}_k^{(3)'} \right)_p \sum_i \left(\boldsymbol{A}_k^{(1)'} \circ \boldsymbol{A}_k^{(2)'} \right)_{ji} \tag{17}$$

不难看出，式（17）就是利用张量分解结果估计原始张量，类似张量补全的工作。不同的是，加入了用户话题相似性约束。在该约束下，对于那些社会影响力大且与话题无关的用户，其影响力的量化数值将分享给大量话题无关且社会影响力小的用户。反映在最终分解结果中的就是在给定话题下，话题无关但是社会影响力大的用户的社会影响力得分将变得相对较小。同理，话题相关的用户将受到那些话题无关用户的影响很小，在张量分解过程中能够很好地保持这些数值的大小。在分解结果中，比起那些话题无关的用户，其用户观点社会影响力得分将变得相对较大，在最终用户观点社会影响力计算中取得较高的分值。因此，在用户相似性的约束下，本文方法最终能够从估计的张量中较好地选出那些话题相关

且社会影响力大的用户。

3　实验结果及分析

3.1　数据描述

如表 1 所示，实验数据来自新浪微博，包括篮球、经济、法律、健康 4 个话题，共 66 754 个用户、282 748 条微博。为了更加详尽地描述数据构成，图 4 统计了所有话题中拥有相同数量级粉丝数的目标用户分布。不难看出，粉丝数量和目标用户数量近似符合幂律分布（在对数—对数坐标下近似为一条直线）。因此该数据中的目标用户具有一定的代表性。

表 1　　实验数据描述

话题	用户间交互关系数量					微博数量
	u_d	u_s	t_{sd}^{+}	$t_{sd}^{\pm}$	t_{sd}^{-}	u_d
篮球	637	34 332	21 190	34 140	8 860	103 767
经济	476	9 841	3 062	7 908	1 932	54 949
法律	486	10 797	3 728	11 319	2 303	60 179
健康	545	9 640	2 826	8 379	979	63 853

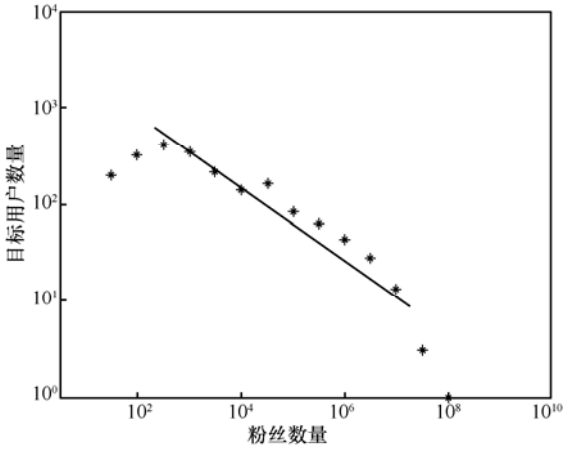


图 4　目标用户与粉丝数量分布

以篮球话题为例，数据内容包含 2 个部分：1) 用户间交互关系；2) 用户信息。其中，用户间交互关系可以使用三元组 (u_s,u_d,t_{sd}) 表示，其中，u_d 表示被评论用户，u_s 表示发表评论的用户，用户 u_s 对用户 u_d 进行了评论并且评论内容是 t_{sd}，t_{sd}^{+}、t_{sd}^{-} 和 $t_{sd}^{\pm}$ 分别表示正面、负面和中性的评论内容。根据预先设定的话题“篮球”，通过新浪微博提供的搜索相关用户功能获取目标用户集合，剩余所需的数据则通过爬取新浪微博页面得到。目标用户将均与篮球相关，即曾发表过与篮球有关的微博，用户间的交互关系是从每个被评论用户各自发表的 40 条微博中获取的。由于评论量可能非常庞大，只选取每条微博的前 30 条评论关系。用户信息则包括用户发表过的微博内容，包括每个被评论用户最多 200 条的微博。

实验的关键是如何确定给定话题下用户观点的社会影响力排序。实验中确定该影响力排序列表的方法将结合用户与话题的相关性，由 5 位均参加过 COAE2013-COAE2015、SIGHAN2015 标注工作的标注者进行标注。提供给这 5 位标注者的数据包括：1）用户列表；2）用户主页地址，可以进入目标用户主页查看该用户的详细情况，包括粉丝数、评论量、职业、发表过的微博等。每位标注者根据这些数据，判断用户在给定话题下的社会影响力大小，然后选出 top5 、top5 ~ top10 和 top10 ~ top20 的用户。如表 2 所示，5 位标注者的 *Kappa* 指标在 0.62 以上，因此对用户观点社会影响力标注在一定程度上是可接受的。

表 2　　数据标注的 *Kappa* 指标

话题	*Kappa* 指标
法律	0.737 574 1
健康	0.620 422 2
经济	0.719 395 0
篮球	0.680 719 3

3.2　实验设计

实验环境为 Matlab 2010，Intel(R) Pentium(R) CPU G645 2.90 GHz，8 GB 内存。将基准方法与本文的方法应用在相同的数据集上，计算得到各个用户在给定话题下的社会影响力得分，即排序结果。最后，基于人工标注的社会影响力用户列表，比较各个方法在不同评价指标的性能优劣。参与实验的基准方法包括以下几方面。

1) CP：未添加本文约束的 CP 分解方法[14]，从分解结果计算用户影响力的方法与本文相同。

2) CP+BM 25：将话题相关性 BM 25 结合 CP 分解方法，计算方法是在 CP 分解的结果上乘以 BM 25 话题相关性得分。

3) OOLAM[11]：OOLAM 模型的计算结果是用户正面影响力和负面影响力 2 个得分，本文对比实验中取正负面影响力的均值作为用户社会影响力得分。

4) OOLAM+BM 25：由于 OOLAM 未考虑话题信息，本文对比实验中将用户话题相关性 BM 25

得分乘以 OOLAM 方法的结果作为用户社会影响力得分。

5) TwitterRank[10]：TwitterRank 的计算结果是用户在特定话题下的重要程度得分，本文实验直接使用该得分作为用户社会影响力得分。

6) TR+RA：由于 TwitterRank 未考虑用户间评论的交互关系。因此在对比实验中，将用户受到评论的数量乘以 TwitterRank 的结果作为用户影响力得分。

7) 分别以用户粉丝量 *FA* 和用户的被评论量 *RA* 度量用户社会影响力。

3.2.1 评价指标

本文所采用的评价指标有以下 3 个指标。

1) 排序精度指标

$$p@k=\frac{\left|A_k \text{ I } B_k\right|}{k} \tag{18}$$

其中，A_k 表示人工排序中的前 k 名用户集，B_k 表示实验排序中的前 k 名用户集。该指标反映了前 k 名实验排序与人工排序的吻合程度。$p@k$ 指标值越大，说明实验得到的排序结果越接近真实情况。

2) 张量分解精度指标

$$RMSE=\sqrt{\frac{\sum_{ijk}(\boldsymbol{X}_{ijk}-\hat{\boldsymbol{X}}_{ijk})^2}{IJK}} \tag{19}$$

其中，X_{ijk} 表示真实数据，$\hat{X}_{ijk}$ 表示预测数据，张量 *RMSE* 的元素个数为 IJK。该指标反映了原始数据与张量分解后的预测数据之间的平均偏差。指标值越小说明张量分解精度越高。

3) 相关性评价指标

使用 Pearson 相关系数来评价本文方法计算的用户社会影响力极性分布与用户真实的社会影响力极性分布的相关强度。计算式如下

$$r=\frac{\sum_{i=1}^{N}(\boldsymbol{X}_i-\bar{X})(\boldsymbol{Y}_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{X}_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(\boldsymbol{Y}_i-\bar{Y})^2}} \tag{20}$$

其中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 表示需要度量相关性的 2 个向量，N 表示这 2 个向量的长度，$\bar{X}$ 和 $\bar{Y}$ 表示均值。实验中，取每个被评价用户收到的正面、中性、负面评价数量作为用户真实的社会影响力极性分布，对这 3 个方面的评价数量做归一化得到 $\boldsymbol{X}$ 的取值。而 $\boldsymbol{Y}$ 的取值就是本文方法对用户社会影响力极性分布的估计值。最后取所有用户的 r，计算均值 $\bar{r}$ 作为评价本文方法反映用户社会影响力极性分布性能的指标。

3.2.2 实验结果分析

1) 参数确定

在确定张量分解中潜在因子维度时，本文根据 *RMSE* 来确定合适的因子维度 R 和迭代次数 it，*RMSE* 越小那么张量分解得到的结果将更能够反映真实数据，从而达到更好的预测效果。

从图 5 可以看出随着 R 的逐渐增大，*RMSE* 的变化趋势趋于平缓，即当 R 的取值到达一定值时，对分解精度的影响将很小。当 R 的取值大于 5 时，*RMSE* 减少的量级仅为 10^{-4}。因此针对该数据集特性，选定 $R=5$ 作为张量分解中潜在因子维度。

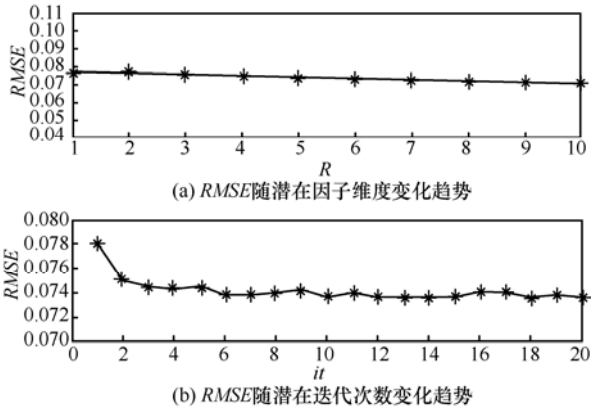


图 5 *RMSE* 随潜在因子维度和迭代次数变化趋势

同样，如图 5 所示，随着迭代次数的增加，*RMSE* 的变化也渐渐趋于缓和，在 it 取值大于 6 时 *RMSE* 的变化量只有 10^{-4} 的量级。因此在本文的实验中，迭代次数 it 取 6。最后根据经验参数 α 取 1，即认为用户相似性约束与张量分解精度相同重要。

2) 用户社会影响力排序精度比较

在 $p@k$ 评价指标下，将本文方法所得到的用户列表与基准方法得到的用户列表进行比较，结果如表 3 所示。从实验结果可以看出，在没有引入本文用户相似性约束以及非负约束时，CP 分解的排序精度较差。其他基准方法在考虑话题信息与用户间交互信息后，性能都有较大提升。本文提出的张量分解方法，通过引入用户相似性矩阵，大大提高了原有 CP 张量分解方法的性能。本文的方法在篮球、经济、法律和健康 4 个话题中 $p@5$、$p@10$ 和 $p@20$ 均值为 0.537 5，比均值第二的 TR+RA 方法性能提升了 12.2%。总体上看，本文的方法相比于其他基准方法在该指标下有较好的表现。

表 3　　本文的方法与基准方法对比实验结果

方法	篮球			经济			法律			健康		
	p@5	*p*@10	*p*@20	*p*@5	*p*@10	*p*@20	*p*@5	*p*@10	*p*@20	*p*@5	*p*@10	*p*@20
CP	0.00	0.00	0.15	0.40	0.30	0.45	0.20	0.20	0.35	0.20	0.20	0.35
CP+BM25	0.00	0.00	0.10	0.40	0.40	0.45	0.20	0.20	0.35	0.20	0.20	0.40
OOLAM	0.20	0.40	0.55	0.60	0.40	0.35	0.20	0.20	0.40	0.20	0.20	0.40
OOLAM+BM25	0.40	0.50	0.55	0.60	0.50	0.55	0.20	0.40	0.55	0.40	0.50	0.55
TwitterRank	0.00	0.20	0.20	0.20	0.10	0.30	0.00	0.10	0.30	0.20	0.30	0.45
TR+RA	0.40	0.50	0.60	0.60	0.50	0.55	0.20	0.40	0.55	0.40	0.50	0.55
RA	0.40	0.40	0.60	0.60	0.40	0.35	0.20	0.30	0.35	0.20	0.20	0.40
FA	0.40	0.50	0.35	0.80	0.70	0.50	0.20	0.60	0.45	0.40	0.30	0.40
本文	0.60	0.60	0.60	0.60	0.60	0.55	0.40	0.30	0.45	0.60	0.50	0.65

3) 用户社会影响力极性特征

为了评价本文方法刻画用户社会影响力极性分布的性能，以用户正面、负面和中性的评论分布作为用户真实的社会影响力极性分布，分别计算每个用户真实社会影响力极性分布与预测结果的 Pearson 相关性得到均值，结果如表 4 所示。篮球、经济、法律和健康这 4 个话题的 Pearson 相关系数值均大于 0.70，具有强相关性。因此本文的方法能够较好地反映用户社会影响力的极性分布。

表 4　　话题的 Pearson 相关系数值

话题	Pearson 相关系数
篮球	0.705 0
经济	0.818 3
法律	0.814 3
健康	0.853 5

根据实验结果，选出一位具有代表性的用户，将其倾向性分布绘图，结果如图 6 所示。该用户的正面社会影响力占主导，可以理解为其他用户对他的反映往往是积极的。不难发现，在本文提出的方法中，借助于用户社会影响力极性分布，可以更加全面的分析用户的社会影响，进而为推荐系统、社交网络信息传播、突发事件检测和广告投放等应用提供更为细致的参考数据。

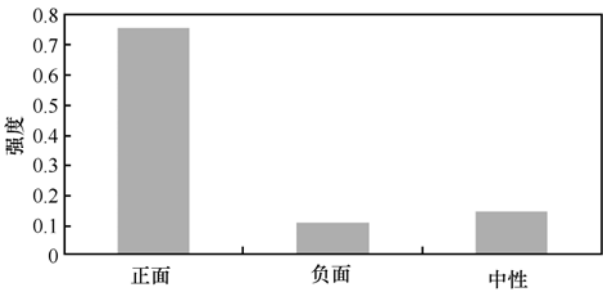


图 6　用户倾向性权重分布

4　结束语

本文提出了一种在给定查询话题下融合用户观点的用户社会影响力分析模型，提出了一种受限的 CANDECOMP/PARAFAC(CP) 分解方法并应用于社会影响力分析。首先，在 CP 分解中加入用户相似性约束，为保证张量分解结果中因子矩阵的可解释性又加入了潜在因子非负约束。其次，为解决受约束的 CP 分解，设计了一种 CP_ALS 的改进算法 HF-CP-ALS 求解本文的模型。最后，通过分析潜在因子评定用户的社会影响力得分，并可以根据张量评论倾向性维度的潜在因子得到用户社会影响力的极性分布，在用户社会影响力的分析上提供了更加详尽的刻画。在与基准方法的对比实验中，本文提出的方法表现出了较好的性能。

参考文献：

[1] CUI P, WANG F, YANG S, et al. Item-level social influence prediction with probabilistic hybrid factor matrix factorization[C]//AAAI. c2011: 331-336.

[2] CUI P, WANG F, LIU S, et al. Who should share what?: item-level social influence prediction for users and posts ranking[C]//The 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, c2011:185-194.

[3] RASHID A M, KARYPIS G, RIEDL J. Influence in ratings-based recommender systems: an algorithm- independent approach[C]//The SIAM International Conference on Data Mining. c2005:556-560.

[4] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: quantifying influence on Twitter[C]//The fourth ACM International Conference on Web Search and Data Mining. ACM, c2011: 65-74.

[5] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]//2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, c2010: 599-608.

[6] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]//The 19th International Conference on World Wide Web. ACM, c2010: 851-860.

[7] BAKSHY E, ECKLES D, YAN R, et al. Social influence in social advertising: evidence from field experiments[C]//The 13th ACM Conference on Electronic Commerce. ACM, c2012: 146-161.

[8] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791-800.
MAO J X, LIU Y Q, ZHANF M, et al. Social influence analysis for micro-blog user based on user behavior[J]. Chinese Journal of Computers, 2014, 37(4): 791-800.
[9] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4):735-752.
WU X D, LI Y, LI L. Influence analysis of online social networks[J]. Chinese Journal of Computers, 2014, 37(4):735-752.
[10] WENG J, LIM E P, JIANG J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//The Third ACM International Conference on Web Search and Data Mining. ACM, c2010: 261-270.
[11] CAI K, BAO S, YANG Z, et al. OOLAM: an opinion oriented link analysis model for influence persona discovery[C]//The fourth ACM International Conference on Web Search and Data Mining. ACM, c2011: 645-654.
[12] KOLDA T G, BADER B W. Tensor decompositions and applications[J]. SIAM Review, 2009, 51(3): 455-500.
[13] DONG Z D, DONG Q. "ZhiHu" [EB/OL]. http://www.keenAge.com.
[14] CICHOCKI A, ZDUNEK R, PHAN A H, et al. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation[M]. John Wiley & Sons, 2009:42-46.
[15] HU X, TANG L, TANG J, et al. Exploiting social relations for sentiment analysis in microblogging[C]//The Sixth ACM International Conference on Web Search and Data Mining. ACM, c2013: 537-546.
[16] DAVIDSON I, GILPIN S, WALKER P B. Behavioral event data and their analysis[J]. Data Mining and Knowledge Discovery, 2012, 25(3): 635-653.
[17] KOLDA T G, BADER B W, KENNY J P. Higher-order Web link analysis using multilinear algebra[C]//Fifth IEEE International Conference on Data Mining. IEEE, c2005: 242-249.

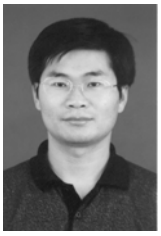
作者简介：

魏晶晶（1984-），女，福建平潭人，福州大学博士生，主要研究方向为网络文本观点挖掘。

陈畅（1991-），男，浙江江山人，福州大学硕士生，主要研究方向为社交网络、数据挖掘等。

廖祥文（1980-），男，福建泉州人，博士，福州大学副教授、硕士生导师，主要研究方向为 Web 信息检索与观点挖掘。

陈国龙（1965-），男，福建莆田人，博士，福州大学教授、博士生导师，主要研究方向为网络计算、智能信息处理等。



程学旗（1971-），男，安徽安庆人，博士，中国科学院计算技术研究所研究员、博士生导师，主要研究方向为网络科学与社会计算、互联网搜索与挖掘等。