ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2018/12(05)-0741-12
doi: 10.3778/j.issn.1673-9418.1709038

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

话题感知下的跨社交网络影响力最大化分析*

任思禹+,申德荣,寇 月,聂铁铮,于 戈
东北大学 计算机科学与工程学院,沈阳 110819

Topic-Aware Influence Maximization Across Social Networks*

REN Siyu+, SHEN Derong, KOU Yue, NIE Tiezheng, YU Ge
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
+ Corresponding author: E-mail: neu_rensy@126.com

REN Siyu, SHEN Derong, KOU Yue, et al. Topic-aware influence maximization across social networks. Journal of Frontiers of Computer Science and Technology, 2018, 12(5): 741-752.

Abstract: With the continuous emergence of various social networking sites, finding a group of the most influential users on multiple social networks is very important for product recommendation or product promotion. In order to improve the breadth and accuracy of product recommendation or promotion, this paper presents an algorithm of topic-aware influence maximization, M-TLTGreedy. Firstly, this paper evaluates the relation among users based on their text semantics and social relationships in multiple social networks to build a topic-based cross-network graph. Then, based on the linear threshold model, this paper designs a topic-aware influence maximization model across social networks, M-TLT (multiple-topic linear threshold) model. Next, this paper uses the improved heuristic algorithm to select a set of users based on the M-TLT model. Finally, the extensive experiments on real datasets show that the M-TLTGreedy algorithm performs well on influence spread and running time.
Key words: social networks; topic-aware; influence maximization; linear threshold model

摘 要:随着各种社交网站的不断涌现,在多社交网络上找到影响传播范围最大的一组用户,对产品推荐或产品推广具有重要作用。为提高产品推荐或推广的广度和精准性,提出了一种跨社交网络基于话题感知的影响力最大化处理方法M-TLTGreedy。首先,根据跨社交网络中的文本语义信息和用户间的社会关系来评价多社交网络中用户间关系,以此构建一个基于话题的跨社交网络图;然后,在线性阈值模型的基础上,设计了一个

* The National Natural Science Foundation of China under Grant Nos. 61472070, 61672142 (国家自然科学基金).
Received 2017-08, Accepted 2017-10.
CNKI网络出版: 2017-10-16, http://kns.cnki.net/kcms/detail/11.5602.TP.20171016.1637.008.html

基于话题感知的跨社交网络影响力最大化模型M-TLT(multiple-topic linear threshold);接着,基于M-TLT模型,利用改进的启发式算法,进行初始用户集的选取;最后,基于大量数据集的实验,证明了该算法无论在影响范围和时间效率上均表现良好。

关键词:社交网络;话题感知;影响力最大化;线性阈值模型

文献标志码:A **中图分类号:**TP391

1 引言

社交网络影响力最大化是指在社交网络上,通过一个影响传播模型,选择 k 个种子用户使其得到最大化预期的影响传播范围。其核心是设计可靠的影响传播模型和有效的搜索算法,能够在有限的预算范围内有效地宣传新产品。

现有的影响力传播研究工作[1-10]主要面向静态的单一社交网络,通过分析社交内用户之间的影响力传播,找出最具有影响力的用户集合。实际上,面向多个社交网络的影响力最大化研究也具有重要意义,因为单一网络可能不能满足广度需求,且单一网络的传播广度不如跨社交网络的传播范围。为此呈现出有关跨社交网络影响力最大化研究[11-13],即在多个社交网络上进行用户的影响力分析。然而,当前研究具有如下局限性:(1)跨社交网络之间通过锚点对进行简单链接,影响传播精度;(2)没有考虑话题因素,制约了种子选择的精准性。有关话题感知的影响力分析研究主要面向单社交网络,但由于不同网络的话题不同,处理多个网络上的话题具有一定难度。

本文研究基于话题感知的跨社交网络的影响力最大化模型M-TLT(multiple-topic linear threshold)。例如,为扩大旅游推广范围,一个旅行社希望在人人网、微博等社交网络上推广一条欧洲行的旅游路线,即在多个社交网络上找到对旅游话题感兴趣的且影响范围最大的一组用户进行推广。本文的主要贡献如下:

(1)提出了一种基于话题的跨社交网络图构建方法。利用用户关注的事物之间的联系,将多个网络联系到一起,并通过用户的话题得到用户之间边的传播概率,最终形成一个基于话题的跨社交网络图。

(2)提出了一个在线性阈值模型下融合话题因素的M-TLT模型,并证明了M-TLT模型的合理性。

(3)设计了一种基于M-TLT模型的改进的启发式种子用户选择算法。首先,基于潜力大小排序来减少候选种子集合及边缘增益的计算量;然后,利用CELF(cost-effective lazy-forward)队列进行种子用户的选择,在选择过程中更新候选种子集合保证最终能够得到 k 个种子用户。

(4)设计了对比实验,证明所提出的模型能够更准确地建模潜在的影响力传播过程,在时间复杂度上优于对比算法。

2 相关工作

Richardson等人[14-15]首先提出了口碑效应(word-of-mouth)概念,研究基于“病毒式营销”的影响力最大化模型。Kempe等人[1]首次把影响最大化问题建模为在传播模型上寻找影响力最大的 k 个节点的离散优化问题。他们提出了两种常用的影响力传播模型:线性阈值模型(linear threshold model,LT)和独立级联模型(independent cascade model,IC)。基于这两个模型的影响力最大化算法,提出了能近似达到最优解(1−1/e)的贪心算法,该算法每一轮选择边缘收益最大的节点,计算代价大,不适用于大规模网络。为了改善贪心算法的低效性,Leskovec等人[2]通过挖掘影响函数的子模特性,提出了CELF算法,减少了种子影响范围次数。进一步,Goyal等人[3]提出了CELF算法的优化算法CELF++算法,效率提高了35%~55%,但同样不适用于大规模网络。接下来,提出了很多基于IC和LT模型的高效启发式算法。例如,Chen等人[4]提出了基于IC模型的利用节点局部树状结构近似影响传播的MIA(maximum influence arborescence)算法;Goral等人[5]提出了在LT模型下的SIMPATH算法。SIMPATH算法在运行时间、内存

损耗和影响范围方面都具有很好的性能。上述方法都没有考虑话题因素,制约了种子用户选择的精准性。

Tang等人[6]研究了用户间topic-wise影响强度的问题,阐述了用户通常只能在某一个领域具有很大的影响力。Aslay等人[7]研究了话题感知的影响力最大化问题(topic-aware influence maximization,TIM),对有限数量可能查询进行预先计算,然后使用树基指数(INT)方法建立索引,有效地改善了查询性能。Barbieri等人[8]提出的话题感知模型中,侧重关注用户权威性和主题的兴趣,没有考虑用户-用户的影响,使传播模型的参数数量急剧减少,从而提高效率。Chen等人[9]提出的话题感知影响力最大化查询中,利用MIA模型近似计算影响传播,并对贪心阶段边缘算法提出了多种优化方法,在时间、影响范围上均优于前几种算法。上述方法考虑了话题因素,但仅在单网上进行研究。

Nguyen等人[11]研究了多社交网络上最小花费的影响力问题,提出了一个利用多种耦合方案将多网问题降到单网解决的模型表示。随后文献[12]在之前研究的基础上进行改进,提出了星式耦合网络,通过减少额外的顶点和边来提高性能。李国良等人[13]研究了多社交网络上的影响力最大化问题,通过自传播将多个网络建立联系,首次提出了针对多社交网络上节点对实体的影响计算模型来评估多网下节点间的影响计算问题,并在独立级联影响模型下提出了多种解决方案,具有较好的伸缩性和时间性能。

与已有工作比较,本文工作具有如下优势:(1)利用社交网络中的文本信息将多个社交网络联系起来,并通过用户的话题分布结合用户之间边的权重评价用户之间的影响概率,使其更加具有可靠性;(2)跨社交网络传播模型中融合了话题因素,提高了种子用户选择的精准性;(3)利用3个优化策略减少了种子用户选择的时间复杂度。

3 问题定义

下面主要介绍本文对跨社交网络影响力最大化的定义。表1为本文使用的符号。

影响力最大化问题是获得网络中信息的传播过程,并以最大限度提高网络中信息的传播范围。首先,给出相关概念;之后,给出基于话题感知的跨社交网络影响力最大化的问题描述。

Table 1 Symbols used in this paper

表1 本文使用的符号

符号	定义
$G^i=(V,F,\omega)(i=1,2,\cdots,m)$	社交图
ξ	用户阈值
$V(a)$	活跃用户集合
Num	匹配对个数
ε	话题相似度阈值
D	文本信息
θ	用户话题分布
$InN(u)$	用户的内向邻居集
$OutN(u)$	用户的外向邻居集
$\Gamma(S\|\boldsymbol{\gamma})$	用户传播范围

定义1(种子用户) 用户 $v\in V$ 作为社交图中 G^i 影响传播的来源,被称为一个种子用户。种子用户集合由 S 表示。

定义2(活跃用户) 如果用户是种子用户或者在影响力传播模型下从已活跃的用户 $v\in V(a)$ 接收了信息,则称为活跃用户。用户一旦变为活跃状态则添加到 $V(a)$ 中。

定义3(传播范围) 给出一个影响力传播模型,种子用户集合的影响传播范围用 S 最终能影响到的用户数量表示,记为 $\Gamma(S)$。

基于话题感知的影响力最大化问题在文献[13]中被提出。例如,一个用户的话题分布为<乒乓球:0.5,游泳:0.3,电子:0.7>,假设人们想推广乒乓球和游泳,那么给出一个话题查询向量(<0.4,0.6,0>,2),即要找到两个种子节点,乒乓球话题占查询40%比重,游泳占查询60%比重,且影响范围最大。

定义4(基于话题感知的影响力最大化) 给出一个话题查询 $Q=(\boldsymbol{\gamma},k)$,基于话题感知的影响力最大化是在社交网络上找到 k 个种子用户集合 S,使集合 S 在此查询下的 $\Gamma(S)$ 最大。

问题1(基于话题感知的跨社交网络影响力最大化) 给出 m 个基于用户话题分布的社交图 $G^i(i=1,2,\cdots,m)$,并给出一个TIM查询 $Q=(\boldsymbol{\gamma},k)$,找到一组种子用户集合,使其对网络上其他用户的总影响力最大,即 $\Gamma(S|\boldsymbol{\gamma})$ 最大。

4 基于话题感知的M-TLT模型

4.1 M-TLT传播模型

M-TLT传播模型核心包括跨社交网络构建、影响概率计算和影响力传播过程三部分。

4.1.1 跨社交网络构建

不失一般性,以两个社交网络为例,介绍本文的模型。

跨社交网络构建的核心是如何模型化社交网络间的关联关系。给定网络 G^1、G^2 的同一实体1和实体2(同一个实体指不同网络上的同一个用户),如果实体1将信息从 G^1 传到 G^2,需要考虑同一实体的邻居用户在跨网络间的某种联系。例如,首先在文本信息中抽取或自定义like匹配对,如<歌手,歌曲>、<演员,电影>等。如图1所示,情况1:G^1 中实体1 follow的用户1 like的歌手/演员与 G^2 中实体2 like的歌曲/电影相对应,说明实体的follow用户可能不通过 G^1 上的用户与 G^2 的影响链接产生影响,而形成一条跨网影响的单向链接,即图中influence链接。情况2:两个实体的follower用户2和用户3同时喜欢同一个匹配对,那么当跨网实体用户传递信息时,其follower之间产生间接的影响,则形成相似similarity的双向链接。这些链接的话题分布与对应实体共同喜欢的话题有关。

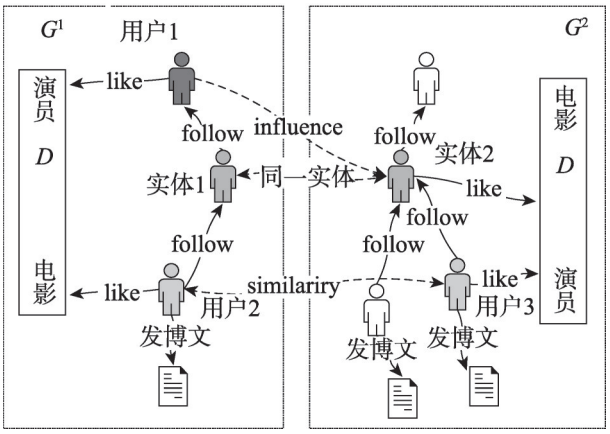

Fig.1 Cross-network connection

图1 跨社交网络连接

对于跨社交网络之间没有直接联系的两个用户,若兴趣相似,他们之间可以产生间接的联系,那么similarity链接具有合理性。因此需要计算在不同的话题查询下两个用户的话题相似度 $sim(\theta_u,\theta_v|\boldsymbol{\gamma})$,当 $sim(\theta_u,\theta_v|\boldsymbol{\gamma})>\varepsilon$ 时,建立用户的"链接"边。本文利用余弦相似度计算用户话题的相似度,见式(1):

$$sim(\theta_u,\theta_v|\boldsymbol{\gamma})=\frac{\theta_u\cdot\theta_v}{\sqrt{\theta_u^2}\sqrt{\theta_v^2}}\cdot\boldsymbol{\gamma} \tag{1}$$

基于上述思想,跨社交网络构建过程见算法1。

算法1 跨社交网络构建算法newTopicGraph

输入:$G^i=(V,F,\omega,\xi)$, Num, θ, ε, D,话题查询 $Q=(\boldsymbol{\gamma},k)$,网络之间实体匹配对 $Alignment$。

输出:跨社交网络 G。

```
1. for every u∈G^i do
2.   for every v∈G^i do
3.     if ( (u,v)∈Alignment ) do
4.       在 u、v 之间建立边;
5.       获取 u、v 的内向邻居集和外向邻居集 InN(u)、InN(v)、OutN(u)、OutN(v)
6.       /*建立influence边*/
7.       for every i∈OutN(u) do
8.         建立 i 到 v 的单向influence边;
9.       end for
10.      for every i∈OutN(v) do
11.        建立 j 到 u 的单向influence边;
12.      end for
13.      /*建立similarity边*/
14.      for every i∈InN(u) do
15.        for every j∈InN(v) do
16.          计算 i 和 j 的话题相似度 sim(i,j);
17.          if ( sim(i,j)>ε ) do
18.            建立 i 和 j 间双向similarity边
19.          end if
20.        end for
21.      end for
22.    end if
23.  end for
24. end for
```

其中,第1~5行遍历不同网络中的用户,若两个用户是同一个实体,则生成一个双向链接,得到 u 和 v 的外向邻居用户集合 $OutN(u)$、$OutN(v)$ 以及内向邻居用户集合 $InN(u)$、$InN(v)$。第7~12行,建立influ-

ence边。判断 $InN(u)$ 是否与 v 有相同的like匹配对，若存在且匹配对的数量大于数值 Num，则建立一个单向的influence链接，其方向指向 v。第13~21行建立similarity边。遍历 $OutN(u)$ 和 $OutN(v)$，若当前话题相似度 $sim(\theta_u,\theta_v|\boldsymbol{\gamma})>\varepsilon$，建立一个双向similarity链接。对不同的链接边标记不同的 $flag$ 值，实体用户 $flag=0$，similarity以及influence链接，$flag=1$。第22~24行，若两个用户不是同一实体，继续遍历网络中其他用户。

4.1.2 链接上的影响概率计算

链接上的影响概率按类型采用不同的计算方法。相同实体之间的传播概率设置为：转发文本的数量除以文本的总数量。而对于网络内的follow以及网络间的influence和similarity链接来说，当计算用户 u 对 v 的影响力时，用户 v 对用户 u 的话题接受度通常不同，因此不能用统一的影响概率来表示，应该对不同的话题有不同的影响概率。利用文献[10]中用户之间的不同话题影响力大小计算公式，如式(2)所示：

$$P^i_{(u,v)}|\theta=\frac{\omega_{(u,v)}\cdot\theta^i_u}{\sum\limits_{j\in InN(v)}(\omega_{(j,v)}\cdot\theta^i_j)},i=1,2,\cdots,z \tag{2}$$

其中，$\omega_{(u,v)}$、$\omega_{(j,v)}$ 是用户之间边权重；θ^i_u 是用户 u 在话题 i 下的话题比重。

4.1.3 影响力传播过程

本文的影响力传播过程基于线性阈值模型。给定 m 个用户话题分布社交图 G^i、初始种子集合 S 和话题查询 $Q=(\boldsymbol{\gamma},k)$，那么影响力传播过程描述如下：在第 t 步，所有在前 $t-1$ 步已经被激活的节点仍然处于活跃状态，任意一个用户 u 接收到的总影响力为 $W^t(u)=\sum\limits_{i=1}^{z}\sum\limits_{v\in InN(u)}\gamma^i\cdot P_{(v,u)}|Z$，若 $W^t(u)\geqslant\xi_u$，即用户收到影响力大于本身阈值，则用户 u 被激活，为活跃状态。持续这个过程直到没有可被激活的节点。

4.2 M-TLT模型性质

如果 $\Gamma(S|\boldsymbol{\gamma})$ 是单调子模函数，那么使用贪心算法进行最大边际增益的近似最优达到 $(1-1/e)$ 结果。

定理1 M-TLT模型下的 $\Gamma(S|\boldsymbol{\gamma})$ 是单调子模函数。

证明 在特定TIM查询 $Q=(\boldsymbol{\gamma},k)$ 下，任意两个用户 u、v 之间的边，均有 $w(u,v)=\sum\limits_{i=1}^{z}\gamma^i\cdot P^i_{(u,v)}|Z$，即对于任意一条边在特定的TIM查询下，都是一个固定值，即特定TIM查询下，用户之间的影响力是确定的。M-TLT模型下的影响力最大化可以看作普通的单网上的影响力最大化问题。根据文献[3]，M-TLT模型下 $\Gamma(S|\boldsymbol{\gamma})$ 是单调子模函数。 □

定理2 M-TLT问题是NP-难问题。

证明 可以用 $z=1$ 的话题分布的条件构建一个M-TLT问题的一个实例，解决M-TLT问题就是解决常规LT模型下的影响力最大化问题。传统影响力最大化问题是NP-难，因此证明了M-TLT问题也是NP-难问题。 □

5 M-TLT模型下种子用户选取算法

由前文讨论可知，可采用贪心算法选择种子用户。首先，需要建立从用户 u 到用户集合 V 中所有的影响模型；然后，迭代计算 S 中每个用户 u 的边缘影响。边缘影响是当前种子集合 S 下，用户 u 加入种子集合与 S 总的影响范围与 S 产生的影响范围之差，计算方式为 $\delta_u(S)=\delta(S\cup\{u\})-\delta(S)$。可见，采用该类贪心算法计算代价大，为此本文采用启发式算法选择种子用户。

本文利用CELF启发算法[4]来优化算法。由于CELF不能保证在最坏情况下减少运行时间，为了有效降低CELF算法的运行时间，本文提出了如下3种优化策略：一是基于节点潜力（见M-Potentiality算法）减少计算代价；二是候选集 C 的选择优化（见M-Candidate算法），因为CELF算法迭代过程中会计算大量不必要计算的节点的边缘增益，所以提出了定理3，大大降低了不必要计算的边缘增益；三是减少边缘增益的计算代价。

5.1 优化策略

5.1.1 基于用户潜力的优化

用户潜力是指用户能够激活其他用户的能力，用户最终可能影响到的节点越多，其潜力越大。用户 u 的潜力评价方法为 u 以及其所有外向邻居边上的影响传播概率和查询话题的乘积之和与用户本身阈值之比，如式(3)所示：

$$Pot_u = \sum_{v \in OutN(u)} \frac{\sum_{i=1}^{z} \gamma^i \cdot P_{(u,v)}|Z}{\xi_v} \quad (3)$$

通过忽略有较小潜在边缘收益的用户,只关注具有较高潜力的用户,来减少节点边缘增益的计算次数。

用户潜力(M-Potentiality)算法见算法2。

算法2 M-Potentiality

输入:$G^i=(V,F,\omega)$,话题查询 $Q=(\boldsymbol{\gamma},k)$,用户阈值 ξ。

输出:按潜力值进行排序的节点集合 P。

1. for all $u \in V$ do

2. 计算 Pot_u;

3. end for;

4. 根据 Pot_u 对节点 u 进行排序;

5.1.2 候选集选择的优化

CELF算法在迭代过程中会计算所有非种子节点的边缘增益,然而被当前种子集合激活节点的边缘增益为0,因此不需要计算其边缘增益。下面给出定理及证明。

定理3 当前种子集合激活的点的边缘增益为0。

证明 如图2所示,假设当前种子集合激活点的边缘增益不为0,即 $\delta_a(S) \geqslant 1$(其中 a 为被当前种子集合 S 激活的任意节点)。那么必然至少存在一个点 q,使得种子集合为 S 时不能激活 q,而当种子集合为 $S \cup \{a\}$ 时能激活节点 q。若当前种子集合为 S,假设在第 t 步激活了节点 a,那么在 t 步之后,根据线性阈值过程[3],节点 a 也将作为活跃节点影响其他节点,即在 t 步之后,a 将与种子集合 S 中的节点一起影响其他节点,记为 $S'=S \cup \{a,\cdots\}$。根据假设当种子集合为 $S \cup \{a\}$ 时能激活节点 q 以及定理1,M-TLT传播模型下的影响力最大化是单调的,因此 S' 也能激活节点 q,这与假设种子集合为 S 时不能激活 q 相矛盾,假设不成立,从而当前种子集合激活节点的边缘增益为0。 □

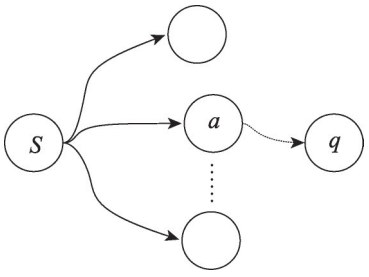

Fig.2 Influence spread of seed set S

图2 种子集合S的影响传播

根据定理3可知,所有被当前种子集合激活节点的边缘增益为0,因此在计算候选节点的边缘增益时,不需要计算已被当前种子集合影响到节点的边缘增益,即可以将被当前种子集合激活的节点从候选集合中删除。

基于算法3获取初始候选集,其中候选集的大小通过参数候选集规模 p 进行确定,返回 $2^p \times k$ 个节点作为候选集。

算法3 M-Candidate

输入:按潜力值排序的节点集合 P,候选集规模 p。

输出:将集合 P 中前 $2^p \times k$ 个元素添加到候选集 C 中。

1. 将集合 P 中前 $2^p \times k$ 个元素添加到候选集 C 中;

2. 将集合 P 中前 $2^p \times k$ 个元素从 P 中删除;

3. 返回候选集 C;

若在当前候选集中选不出 k 个种子用户,需要对候选集合进行扩展。算法4是候选集更新算法,增益节点集合是将当前增益最大节点新激活的节点集合。如果队列 QC 的大小小于 $2^{p-1} \times k$ 个,那么执行第3~5行。新加入的节点采用lazy策略,不计算新加入节点的边缘增益,直接赋值为CELF队列 QC 中用户边缘增益的最小值。

算法4 M-Candidate-Update

输入:CELF队列 QC,增益节点集合 M,按潜力值排序的节点集合 P,候选集合规模 p。

输出:更新后的CELF队列 QC。

1. 将 M 中的节点从 QC 和 P 中删除;

2. if $(|QC| < 2^{p-1} \times k))$;

3. 从集合 P 中取前 $2^{p-1} \times k$ 个元素放到 QC 中;

4. 将取出的 $2^{p-1} \times k$ 个元素从 P 中删除;

5. 返回CELF队列 QC;

5.1.3 边缘增益计算代价的优化

种子集合 S 能够激活节点,那么种子集合 $S \cup \{a\}$ 也一定可以激活。然而在计算节点 a 的边缘增益 $\delta_a(S)$ 时,往往需要重新从 $S \cup \{a\}$ 开始一步一步地迭代计算,导致已经确定能被激活的点被重复计算。

因此使用一个集合 O 来存储当前种子集合能够影响到的节点的集合，那么节点 a 的边缘增益为 $\delta_a(S)=\delta(S\cup O\cup\{a\})-\delta(S)$，从而降低了边缘增益的计算量。

5.2 基于贪心的种子用户选择算法

最后，采用贪心算法M-TLTGreedy算法（见算法5）找到大小为 k 的种子用户集合 S。其中，第1行初始化种子用户集合、CELF队列以及活跃节点集合 M。第2~4行连接多个基于话题的社交网络并计算用户的潜力值和候选种子用户集合。第5~8行计算候选集合中用户的影响力大小并插入到队列 QC 中。若此时种子集合大小小于 k，计算队头用户的影响力大小，直到 u 仍然在CELF队列 QC 的队头（第9~13行）。若没有增益返回种子集合 S，将用户 u 加入种子集合中，更新候选集合，直到得到 k 个种子用户算法停止。

算法5 M-TLTGreedy算法

输入：$G^i=(V,F,\omega)$, D, Num, p, 话题查询 $Q=(\gamma,k)$，用户阈值 ξ。

输出：大小为 k 的种子用户集合 S。

1. 初始化 $S\leftarrow\varnothing$，CELF队列 $QC\leftarrow\varnothing$，活跃节点集合 $M\leftarrow\varnothing$；
2. $G\leftarrow newTopicGraph(G^i,Num,\theta,\varepsilon,D,Q)$；
3. $P\leftarrow M-Potentiality(G^i,Q(\gamma,k))$；
4. $C\leftarrow M-Candidate(P,p)$；
5. for $u\in C$ do
6. 计算 $\delta_u(S)$；
7. 将 u 插入到队列 QC 中；
8. end for
9. while ($|S|<k$) do
10. repeat
11. $u\leftarrow QC.top$；
12. 重新计算 $\delta_u(S)$；
13. until1 u 仍然在CELF队列 QC 的顶端；
14. if ($\delta_u(S)\leqslant 0$) then
15. return S；
16. end if
17. $S\leftarrow S\cup\{u\}$；
18. 获取节点 u 的活跃节点集合 M；
19. $QC\leftarrow M-Candidate\text{-}Update(QC,M,P,p)$；
20. end while

6 实验与结果

下面基于M-TLT模型，在两组数据集（微博-知乎和Twitter-Facebook[16]）上测试了本文提出的算法。

6.1 数据集

本文使用两组数据集，真实的微博-知乎数据集和Twitter-Facebook数据集，具体的统计信息见表2和表3。微博-知乎（Weibo-Zhihu）数据集包括微博以及抓取的知乎信息，其中也包括作为话题提取的文本信息。本文采用STRM（social-relational topic model）[17]算法来计算用户的话题分布。Twitter-Facebook数据集没有文本信息，本文随机生成了用户的话题分布。由于现有的数据集中无法获得用户的阈值信息，在实验中统一设置为0.5。随机生成话题查询。

Table 2 Statistics of Weibo-Zhihu datasets

表2 微博-知乎数据集统计

数据集	用户数	用户关系数	文本数	话题数	匹配数
微博	63 642	1 391 719	84 169	20	452
知乎	232 826	1 814 775	61 123	20	

Table 3 Statistics of Twitter-Facebook datasets

表3 Twitter-Facebook数据集统计

数据集	用户数	用户关系数	匹配数
Twitter	669 198	3 710 789	328 224
Facebook	422 291	12 749 257	

所有的对比实验中各个算法均使用相同的数据集、用户阈值和话题查询，不会对实验结果产生影响。

6.2 实验设置及对比实验

将本文提出的M-TLTGreedy算法与已有的算法进行对比。评价指标：一是影响范围，范围越大表明种子用户的质量越好；二是运行时间，时间越短表明种子用户选择的算法越好。本文比较了几种算法在以上两个数据集上的影响范围和运行时间。

CELF：用CELF[6]进行优化的贪心算法。由于CELF算法不适用于基于话题的跨网影响力计算，将算法1生成的网络图作为单一网络用CELF算法进行计算。

SIMPATH：该算法提出了两个优化，一是利用最大顶点覆盖减少第一次迭代中的调用次数，二是用参

数 l 在迭代开始选择最有前途的候选种子，l 取值4。

HighDegree[3]：该算法把度数高的节点作为有影响力的节点，选出 k 个最高出度的节点作为种子集。

M-TLTGreedy-no3：不用边缘增益优化策略进行优化的M-TLTGreedy算法。

CELF-3：计算边缘增益时，利用本文的边缘增益优化策略进行优化的CELF算法。

CELF-2：用本文的候选集选择优化策略进行优化的CELF算法。

实验环境：Intel® Xeon® CPU E3-1226 v3 @ 3.30 GHz，内存16 GB；操作系统：CentOS Linux release 7.3.1611 (Core)。

6.3 实验和结果

以下通过4个实验来验证本文算法的性能。前3个实验都是在固定候选集规模 $p=4$ 条件下进行。

6.3.1 优化策略比较

(1)用户潜力优化策略的作用

用户潜力优化策略旨在寻找有较高可能成为种子节点的用户。只计算具有较高潜力用户的边缘收益，从而降低了CELF算法第一次迭代过程中的计算量。实验比较了M-TLTGreedy算法和CELF算法在第一次迭代中的运行时间，其中M-TLTGreedy算法参数为 $p=4$，$k=100$，实验结果如表4所示。可以看出，因为M-TLTGreedy算法只计算部分节点的边缘收益，所以运行时间要明显小于CELF算法。

Table 4 Running time comparison of potentiality optimization strategy

表4 用户潜力优化策略运行时间比较

数据集	算法	运行时间/s
微博-知乎	M-TLTGreedy	17.87
	CELF	187.89
Twitter-Facebook	M-TLTGreedy	43.30
	CELF	570.81

(2)候选集选择优化策略的作用

候选集选择优化旨在降低计算边缘增益节点的数量。实验比较了CELF算法和CELF-2算法的运行时间，其中微博-知乎数据集的用户阈值设置为0.4，Twitter-Facebook数据集的用户阈值设置为0.3。从图3中可以看出，随着 k 的增加，候选集优化的效果越来越明显。

(3)边缘增益优化策略的作用

实验比较了M-TLTGreedy和M-TLTGreedy-no3、CELF-3和CELF算法的运行时间。从图4中可以看出，在两个数据集下，边缘增益计算优化策略均使两个算法的运行时间降低，且随着 k 值的增大，优化效果越明显。这是因为随着种子节点个数的增加，重复计算量也会增大。

6.3.2 影响范围比较

实验比较了M-TLTGreedy、CELF、HighDegree、SIMPATH算法在种子集合大小为10~100时的影响

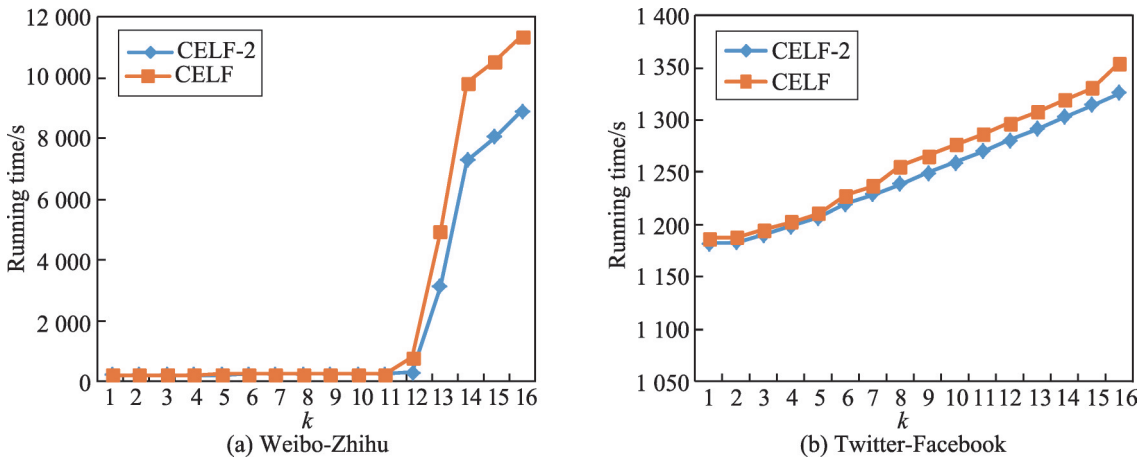

Fig.3 Running time comparison of candidate optimization strategy

图3 候选集选择优化运行时间比较

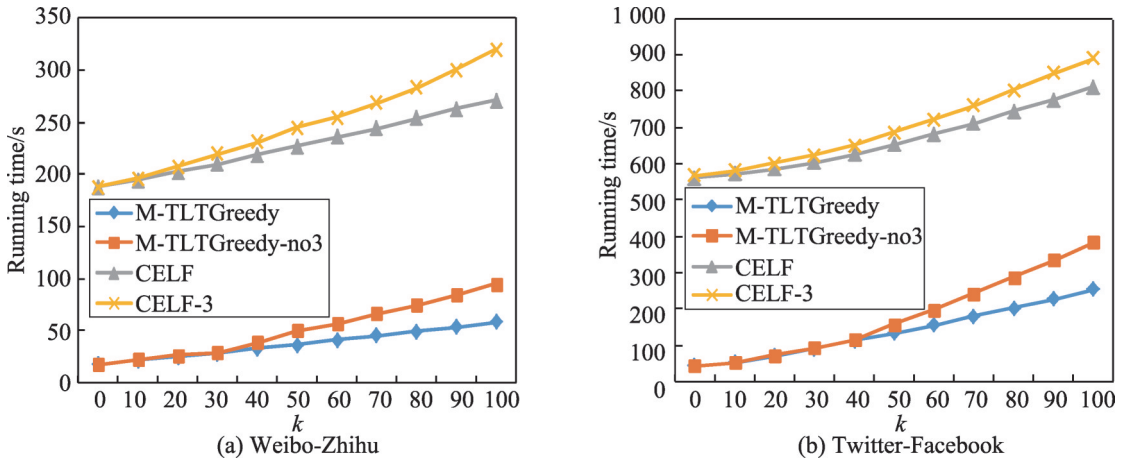

Fig.4 Running time comparison of margin gain optimization strategy

图4 边缘增益优化策略运行时间比较

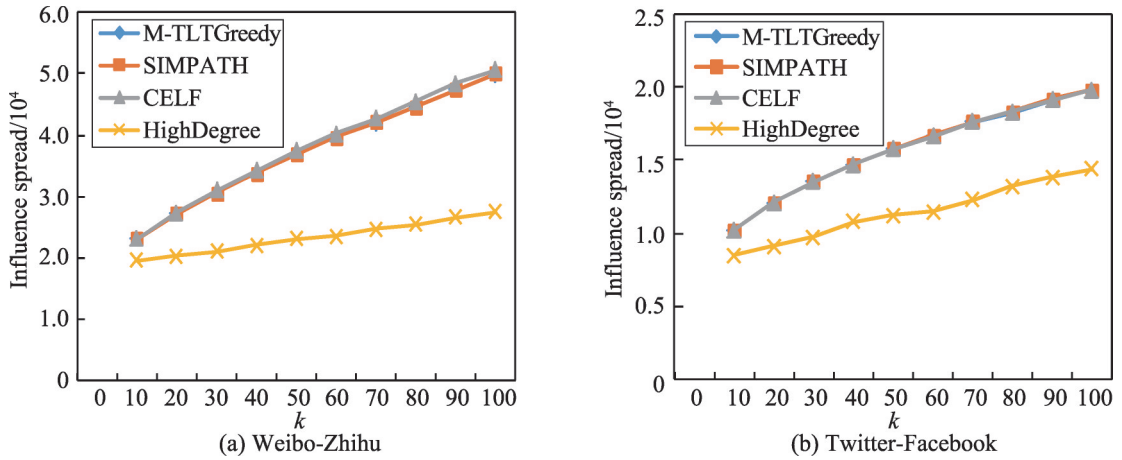

Fig.5 Influence spread comparison

图5 影响范围比较

范围。从图5中可以看出:在微博-知乎和Twitter-Facebook两个数据集上,在任意种子集合大小下,M-TLTGreedy算法的影响范围趋近于CELF、SIMPATH算法的影响范围,明显优于HighDegree算法的影响范围,说明本文算法保证了精度。

6.3.3 运行时间比较

实验比较了M-TLTGreedy、CELF、SIMPATH、High-Degree算法在种子集合为0~100时的运行时间。从图6中可以看出:因为HighDegree算法不需要计算节点的边缘增益,只需要数据集中每个节点的出度,所以HighDegree算法的运行时间非常短;M-TLT-Greedy算法要明显低于CELF算法的运行时间,因为CELF算法在计算增益过程中需要计算图中所有节点用户的增益,并在之后的计算过程中计算了一些不必计算的节点,而M-TLTGreedy算法通过计算用户潜力,得到一个候选种子集合,只对候选集中的节点进行增益计算,同时通过边缘增益优化算法减少了重复计算,所以大大减少了计算时间。SIMPATH算法和M-TLTGreedy算法都针对贪心算法的问题进行了改进。可以看出:在k值小的情况下本文算法所花费的时间更短,随着k值的增大,SIMPATH算法的运行时间稍微优于本文算法。

6.3.4 候选集规模的影响

测试候选集规模p取值对M-TLTGreedy算法影

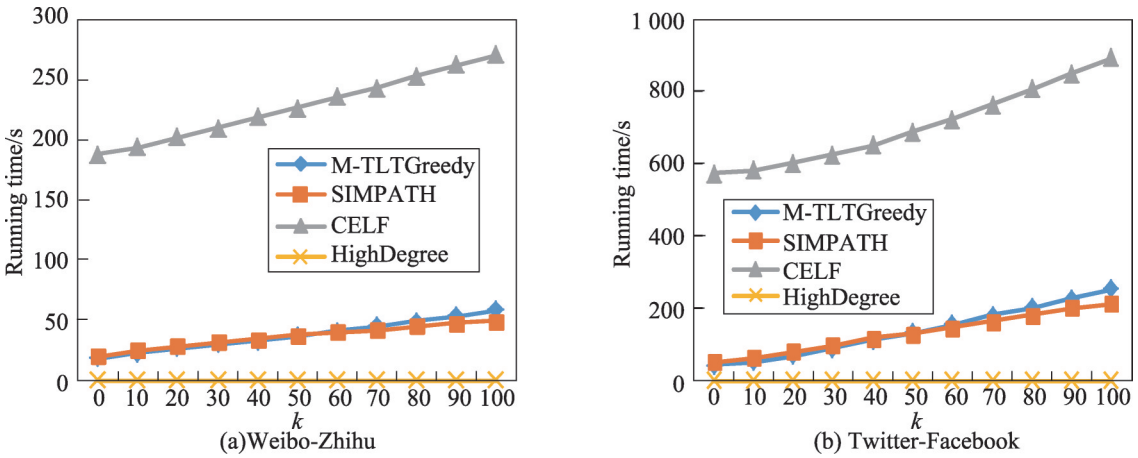

Fig.6 Running time comparison

图6 运行时间比较

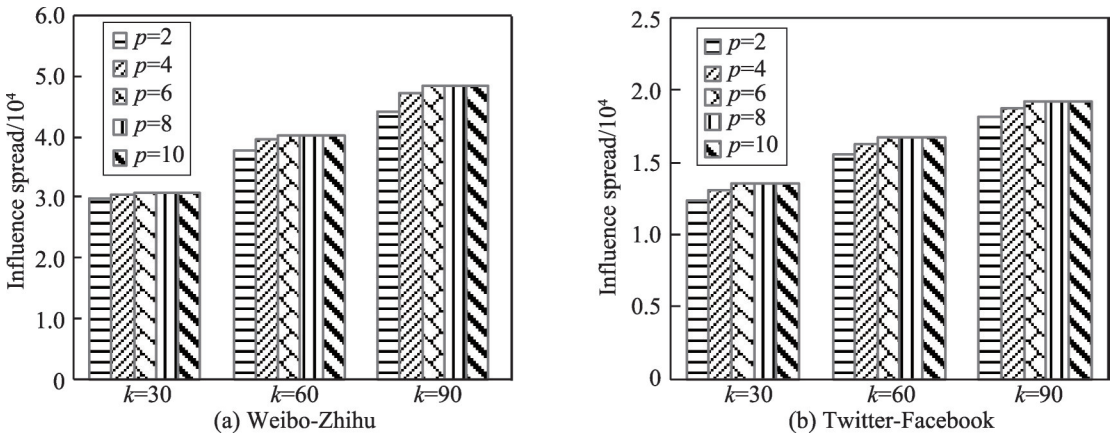

Fig.7 Candidate scale vs. influence spread

图7 候选集规模-影响范围

响范围以及运行时间的影响。图7显示了 $p=2,4,6,8,10$ 时，在两个数据集下的影响范围大小的对比。显然，固定种子集大小，在微博-知乎数据集以及Twitter-Facebook数据集上 $p=2,4,6$ 影响范围有小幅度增加，$p=6$ 之后影响范围基本相同，说明候选集规模对种子用户影响范围的影响不大。

图8显示了分别在不同种子集合大小（$k=30,60,90$）下，随着候选集规模 p 的增大，运算时间也会大幅增加。这是因为随着 p 值的增大，候选集规模也在增大，需要计算的边缘增益的节点数也会大幅增加。

总体看本文提出的M-TLTGreedy算法表现出较高的性能，能够得到一定的影响范围，并在运行时间上优于对比算法。

7 结束语

本文利用用户关注的事物之间的联系，将多个网络联系到一起，并进行影响力模型的建模，提出了M-TLT模型。设计了在M-TLT模型上寻找种子用户集的算法，从而找到跨社交网络上在话题下的影响力最大的 k 个种子用户，解决了对于跨社交网络的话题查询下找到影响力最大的用户集合需求。本文提出的影响力最大化方法主要是针对跨社交网络基于话题的查找，提出的M-TLTGreedy算法实验结果表明，本文算法在影响范围和运行时间方面都能达到满意的效果。下一步，将本文的影响力传播模型扩展到动态的多个网络上，使其更符合现实情况，并设计出更高效的种子选择方法。

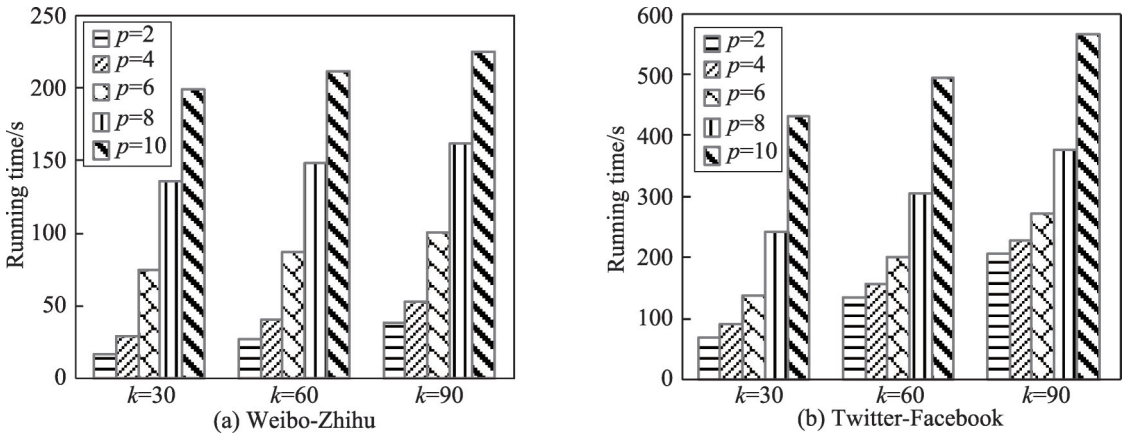

Fig.8 Candidate scale vs. running time

图8 候选集规模-运行时间

References:

[1] Kempe D, Kleinberg J M, Tardos É. Maximizing the spread of influence through a social network[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Aug 24-27, 2003. New York: ACM, 2003: 137-146.

[2] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, Aug 12-15, 2007. New York: ACM, 2007: 420-429.

[3] Goyal A, Lu Wei, Lakshmanan L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]//Proceedings of the 20th International Conference on World Wide Web, Hyderabad, Mar 28-Apr 1, 2011. New York: ACM, 2011: 47-48.

[4] Chen Wei, Wang Chi, Wang Yajun. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Jul 25-28, 2010. New York: ACM, 2010: 1029-1038.

[5] Goyal A, Lu Wei, Lakshmanan L V S. SIMPATH: an efficient algorithm for influence maximization under the linear threshold model[C]//Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, Dec 11-14, 2011. Washington: IEEE Computer Society, 2011: 211-220.

[6] Tang Jie, Sun Jimeng, Wang Chi, et al. Social influence analysis in large-scale networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, Jun 28-Jul 1, 2009. New York: ACM, 2009: 807-816.

[7] Aslay Ç, Barbieri N, Bonchi F, et al. Online topic-aware influence maximization queries[C]//Proceedings of the 17th International Conference on Extending Database Technology, Athens, Mar 24-28, 2014: 295-306.

[8] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models[C]//Proceedings of the 12th International Conference on Data Mining, Brussels, Dec 10-13, 2012. Washington: IEEE Computer Society, 2012: 81-89.

[9] Chen Shou, Fan Ju, Li Guoliang, et al. Online topic-aware influence maximization[C]//Proceedings of the 41st International Conference on Very Large Data Bases, Kohala Coast, Aug 31-Sep 4, 2015: 666-677.

[10] Chu Yaping, Zhao Xianghui, Liu Sitong, et al. An efficient method for topic-aware influence maximization[C]//LNCS 8709: Proceedings of the 16th Asia-Pacific Web Conference on Web Technologies and Applications, Changsha, Sep 5-7, 2014. Berlin, Heidelberg: Springer, 2014: 584-592.

[11] Nguyen D T, Zhang Huiyuan, Das S, et al. Least cost influence in multiplex social networks: model representation and analysis[C]//Proceedings of the 13th International Conference on Data Mining, Dallas, Dec 7-10, 2013. Washington: IEEE Computer Society, 2013: 567-576.

[12] Zhang Huiyuan, Nguyen D T, Zhang Huiling, et al. Least cost influence maximization across multiple social networks [J]. IEEE/ACM Transactions on Networking, 2016, 24(2):

929-939.

[13] Li Guoliang, Chu Yaping, Feng Jianhua, et al. Influence maximization on multiple social networks[J]. Chinese Journal of Computers, 2016, 39(4): 643-656.

[14] Brown J J, Reinegen P H. Social ties and word-of-mouth referral behavior[J]. Journal of Consumer Research, 1987, 14 (3): 350-362.

[15] Richardson M, Domingos P M. Mining knowledge-sharing sites for viral marketing[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Jul 23-26, 2002. New York: ACM, 2002: 61-70.

[16] Cao Xuezhi, Yu Yong. ASNets: a benchmark dataset of aligned social networks for cross-platform user modeling[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, Oct 24-28, 2016. New York: ACM, 2016: 1881-1884.

[17] Guo Weiyu, Wu Shu, Wang Liang, et al. Social-relational topic model for social networks[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Oct 19-23, 2015. New York: ACM, 2015: 1731-1734.

附中文参考文献:

[13] 李国良, 楚娅萍, 冯建华, 等. 多社交网络的影响力最大化分析[J]. 计算机学报, 2016, 39(4): 643-656.

REN Siyu was born in 1992. She is an M.S. candidate at School of Computer Science and Engineering, Northeastern University. Her research interest is influence maximization.

任思禹(1992—),女,辽宁海城人,东北大学计算机科学与工程学院硕士研究生,主要研究领域为影响力最大化。

SHEN Derong was born in 1964. She received the Ph.D. degree in computer software and theory from Northeastern University in 2004. Now she is a professor and Ph.D. supervisor at Northeastern University, and the senior member of CCF. Her research interests include Web data processing and distributed database.

申德荣(1964—),女,辽宁沈阳人,2004年于东北大学计算机软件与理论专业获得博士学位,现为东北大学教授、博士生导师,CCF高级会员,主要研究领域为Web数据处理,分布式数据库。

KOU Yue was born in 1980. She received the Ph.D. degree in computer software and theory from Northeastern University in 2009. Now she is an associate professor at Northeastern University, and the member of CCF. Her research interests include entity resolution and Web data management.

寇月(1980—),女,辽宁沈阳人,2009年于东北大学计算机软件与理论专业获得博士学位,现为东北大学副教授,CCF会员,主要研究领域为实体识别,Web数据管理。

NIE Tiezheng was born in 1980. He is an associate professor at Northeastern University, and the member of CCF. His research interests include data quality and data integration.

聂铁铮(1980—),男,辽宁沈阳人,博士,东北大学副教授,CCF会员,主要研究领域为数据质量,数据集成。

YU Ge was born in 1962. He is a professor and Ph.D. supervisor at Northeastern University, and the senior member of CCF. His research interests include data stream, data mining and distributed database.

于戈(1962—),男,博士,东北大学教授、博士生导师,CCF高级会员,主要研究领域为数据流,数据挖掘,分布式数据库。