

意见领袖识别中的文本倾向性研究*

樊兴华, 吴昊

(重庆邮电大学 中文信息处理研究所 计算智能重庆市重点实验室, 重庆 400065)

摘要: 由于网络用户讨论的主题变化频繁, 因此在进行倾向性判定时, 难以预先构造出满足各种情况的训练语料。针对上述问题, 提出了一种意见领袖识别中的文本倾向性判定方法, 进而建立考虑回复帖子倾向性的意见领袖发现模型。该模型建立在影响力扩散概率模型(IDPM)上, 模型中引入了考虑帖子倾向性的有效系数。实验表明, 该方法是有效的, 其前50个的平均精确率相对分别提高了10.97%和5.45%。

关键词: 意见领袖识别; 倾向性分析; 影响力扩散概率模型

中图分类号: TP301 **文献标志码:** A **文章编号:** 1001-3695(2013)09-2613-03

doi:10.3969/j.issn.1001-3695.2013.09.013

Research of text orientation in opinion leaders identification

FAN Xing-hua, WU Hao

(Chongqing Key Laboratory of Computational Intelligence, Institute of Chinese Information Processing, Chongqing University of Posts & Telecommunications, Chongqing 400065, China)

Abstract: Because the topics changed rapidly which network users discussed, it was difficult to pre-construct the training corpus to meet a variety of situations. To solve the problem above, this paper proposed a method to analyze the orientation in the opinion leaders identification, then established an opinion leader identification model which was based on the influence diffusion probability model (IDPM) and introducing the effective coefficient considering the orientation of the replied comments. The experiments show that the proposed method is effective, the average precisions of the top 50 numbers increased by 10.97% and 5.45% respectively.

Key words: opinion leader identification; orientation identification; IDPM

0 引言

随着互联网对社会生活的逐渐渗透, 人们越来越习惯于利用新兴媒体, 如微博客、QQ、MSN、聊天室和论坛等平台交流心得、参与公众话题讨论。在信息传播的过程中, 意见领袖作为一种重要力量, 对社会舆论的形成发挥着不可忽视的作用。局部意见在意见领袖的引导下演化为舆论, 影响力直接渗透到现实社会^[1,2]。网络意见领袖是指那些通过在新兴媒体发表帖子(文本), 或者回复其他网络用户发表的帖子这种基于文本的交流方式, 将自己的见解、观点传递给其他网络用户, 引起他们内心的共鸣, 进而影响、改变他们的观点、思想和决策的网络用户。与传统意见领袖相比, 网络意见领袖在概念上有着本质的区别, 他们的一些社会特征, 如学位、职位等信息被隐藏。

目前国内外学者对网络意见领袖的识别方法进行了广泛的研究。陈然^[1]对现有的方法进行了较好的梳理和详细分析, 主要识别方法有自我报告法、简单统计测量方法、网络分析方法、运用 PageRank 思想的识别方法、影响力扩散模型(IDM)。樊兴华等人^[3]根据影响力扩散模型存在的缺陷提出了影响力扩散概率模型(IDPM), 利用 IDPM 识别意见领袖时不仅解决了 IDM 中的内容断层和结构断层的问题, 而且引入整合诸如帖子倾向性、用户统计信息、帖子回复结构网络特性

等因素的有效因子。上述方法中, 一方面有的方法并没有考虑帖子倾向性对意见领袖识别的影响, 另一方面有的方法即使考虑了帖子倾向性, 但并未给出有效因子的具体算法和回复帖子倾向性判定方法。

文本倾向性分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。通过对这些具有主观性的文本进行倾向性分析, 能够了解用户对某一事件或事物的看法和观点^[4-6]。而回复帖子的倾向性直接影响了该用户是否为真正的意见领袖。如果大多数的回复帖子对楼主持有反对态度, 那么显然该楼主并非真正的意见领袖; 但倘若在回复帖子中大部分为支持和赞同的响应, 那么该楼主才是真正的意见领袖。可见, 意见领袖识别中应该考虑帖子的倾向性。而意见领袖识别中的文本倾向性判定与常规的文本倾向性判定是不同的。由于前者中网络用户谈论的主题变化非常频繁, 例如今天是有关南海争端的话题, 明天则可能是与官员腐败相关的话题, 因此很难预先构造出满足各种情况的训练语料。因此, 直接将常规的文本倾向性判定方法运用于意见领袖识别中是不合适的。

本文的目的就是针对上述存在的问题, 提出一种意见领袖识别中的文本倾向性判定方法, 进而建立考虑回复帖子倾向性的意见领袖发现模型。该模型建立在 IDPM 的基础上, 并引入了考虑帖子倾向性的有效因子。

收稿日期: 2012-12-22; **修回日期:** 2013-02-05 **基金项目:** 国家自然科学基金资助项目(60703010); 重庆市自然科学基金计划项目(CSTC, 2009BB2079)

作者简介: 樊兴华(1972-), 男, 重庆人, 教授, 博士, 主要研究方向为人工智能、自然语言处理、信息检索(fanxh@cqupt.edu.cn); 吴昊(1985-), 男, 广西玉林人, 硕士, 主要研究方向为中文信息处理。

1 意见领袖发现概率模型

1.1 IDPM

用户帖子回复结构(URS)定义为:由用户发帖导出的、通过回复关系链接在一起的帖子及其回复关系的集合,如图1所示。

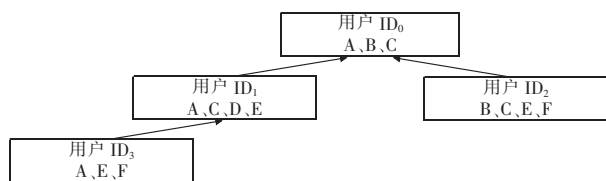


图1 用户帖子回复结构示例

图1中,用户ID₀发表了主帖A、B、C,用户ID₁和ID₂分别回复了帖子A、C、D、E和帖子B、C、E、F,用户ID₃回复ID₁帖子A、E、F,它们形成了一棵回复结构树。

用户所发帖子X的影响力 INF_X :帖子中被有效传递关键词的影响传播之和。假设帖子的关键词向量为 $\langle term_1, term_2, \dots, term_n \rangle$,INF计算公式如下:

$$INF_X = \sum_{i=1}^n w_i \lg P_{IDP}(term_i) \quad (1)$$

其中: $P_{IDP}(term_i)$ 表示关键词 $term_i$ 在相同兴趣空间G内的影响传播概率,即 $term_i$ 在G内传播频率之和与所有term在G内传播频率之和的比值; w_i 为用户所发帖子X的有效传递因子。帖子关键词term的所谓有效传播是指它在以该帖子X为根的帖子回复结构中至少再出现一次。 w_i 定义如下:

$$w_i = \begin{cases} 1 & term_i \text{ 为有效关键词} \\ 0 & \text{其他} \end{cases}$$

1.2 意见领袖筛选模型

一个网络用户可能发表多个帖子,这些帖子中既有主帖,也有对其他用户的回复帖,甚至还有一些灌水帖。有些帖子可能受到大家的热捧,而有些帖子观点则会受到其他用户的批判。除去基于传播内容的影响之外,帖子的其他因素也可能会对意见领袖识别有影响,如帖子的倾向性等。

用户的影响力 $user_{INF}$ 定义为:用户所发全部有效帖子的影响力之和。假设用户发表了m个帖子, $user_{INF}$ 计算公式为

$$user_{INF} = \sum_{i=1}^m u_i INF(X_i) \quad (2)$$

其中: u_i 为帖子有效系数,用于整合诸如帖子倾向性、用户统计信息、帖子回复结构网络特性等因素对用户最终影响力计算的印象。

2 考虑文本倾向性的意见领袖发现模型

回复帖子倾向性与意见领袖的发现有着密切的联系。为此本文引进了考虑帖子倾向性因子,使某一网络用户的影响力得分更接近实际的情况,从而更加有效地识别出真正的意见领袖。

2.1 构造意见领袖识别中的文本倾向性判定训练集

为了得到意见领袖识别中的文本倾向性判定训练集,本文作出如下假设:利用IDPM计算得到的意见领袖排名靠前的网络用户的评论均为正倾向类评论,排名靠后的网络用户的评论为其他类别评论。

具体构造方法如下:首先利用IDPM计算得到各个用户的得分,从而得到一个意见领袖排名;然后选取排名靠前的n个

用户的评论作为训练集的正倾向类评论,选取排名靠后的m个用户的评论为训练集的其他类别评论,进而构造一个意见领袖识别相关的文本倾向性判定训练集。在选择n和m值时,应该进行适当比例的选择抽取。

2.2 帖子倾向性判断

2.2.1 特征扩展

文本倾向性分类与一般的主题分类不同,倾向性文本中具有语义倾向性或情感倾向性的词语或短语对分类起到至关重要的作用。本文利用三步特征扩展的方法^[7]对数据集进行重构,即利用倾向性词表、否定词表、程度副词表对数据集进行特征扩展,把具有语义倾向性的词语或短语加到特征序列中,从而增强特征项对倾向性文本的表述能力。

2.2.2 分类器构造

1) 特征选择

常用的特征选择方法有文档频率方法、互信息方法、信息增益方法、期望交叉熵方法、卡方统计方法等。而对于文本倾向性分类来说,在考虑特征选择方法时,既要考虑具体的语料库特点,又要考虑文本倾向性的特殊性。网络评论语料具有词语传播的重复度高、倾向性强等特点。因此,本文针对网络评论语料的特点提出了考虑词语倾向性的文档频率特征选择公式。

$$DF_Sen(t, c) = \frac{\lg(DF_t)}{\lg(N_c)} \times \frac{\alpha_t(|\beta_t| + 1)}{\alpha_t + \gamma} \quad (3)$$

其中: DF_t 是特征t在类c中出现的文档数; N_c 是c类中的文档总数; β_t 表示特征t的倾向性强度值; α_t 表示特征t所包含的词语的个数; γ 为可通过实验调节的权衡系数。在进行参数选择时,设置一个阈值 DF_SEN_{min} ,把阈值小于某个值的特征项删除,本文通过实验选取的阈值是0.04。当某个特征词均在多类中出现时,则根据该特征词在各类别中的特征得分进行选择,即将在各类别中出现频率相近的特征项删除,出现频率相差较大则保留。本文通过实验选取的是两类中特征得分之差的绝对值大于0.12的词语作为特征。

2) 分类过程

具体的分类器过程:本文利用两步分类的方法^[8]构造两个分类器 CF_1 和 CF_2 ,它们均是利用式(3)进行特征选择的。在分类时,先使用分类器 CF_1 对评论文本进行分类,对分类结果中的可靠部分直接作出判定,而对分类结果中的不可靠部分利用分类器 CF_2 进行二次分类并作出判定,最后得到所有回复帖子的倾向性。

(1) 构造分类器 CF_1

分类器 CF_1 为朴素贝叶斯分类器。将文本d表示为 $d = (t_1, t_2, \dots, t_n)$,其中 t_k 表示文本的特征项。那么朴素贝叶斯分类器表示为

$$P(C_i | d) = \max(P(d | C_i)P(C_i)) \quad i = 1, 2 \quad (4)$$

其中: $P(d | C_i) = \prod_{j=1}^n P(t_j | C_i)$ 。

(2) 构造分类器 CF_2

分类器 CF_2 采用特征倾向性和特征条件概率相结合的重计算方法:

$$f(d) = \sum_{i=1}^n \left(\frac{\lg(P(t_k | C_1)/P(t_k | C_2))}{\lg(P(t_k | C_1)P(t_k | C_2))} \times \lg \left(\frac{P(t_k | C_1)}{P(t_k | C_2)} \right) + \lambda Q(t_k) \right) \quad (5)$$

其中: c_1 表示正向评论类; c_2 表示其他类别评论类; $Q(t_k)$ 为特征 t_k 在倾向性词集中的倾向强度值(如果是正向情感词则 $Q(t_k) > 0$;如果是负向情感词则 $Q(t_k) < 0$;如果特征 t_k 不在情

感词表中,则 $Q(t_k) = 0$; λ 为根据实验确定的调整系数。如果 $f(d) > 0$ 则判定文本 d 为正向评论,否则判定文本 d 为其他类别评论。

2.3 考虑文本倾向性的意见领袖识别模型

利用 2.2 节所得的结果计算考虑帖子倾向性的有效系数,即在一个话题树中,具有正倾向性的回复帖子数在总的回复帖子数中所占的比例。该有效系数可以用如下公式表示:

$$u_i = \frac{\text{一个话题树中判断为正倾向性的回复帖子数}}{\text{该话题树中回复帖子总数}} \quad (6)$$

不同的话题, u_i 的值是不一样的。 u_i 越大,表明支持该主帖的人数越多,其成为网络意见领袖的可能性也就越大; u_i 越小,表明反对该主帖的人数越多,其成为网络意见领袖的可能性就会降低。计算得到该有效系数后,再根据 1.2 节中的式(2)计算得到各个用户的最终影响力得分,得到意见领袖排名。

3 实验

3.1 实验数据

本文使用的数据集是从门户网站网易的社会版块上收集的楼主主题及其评论数据。

利用文献[3]中的 IDPM 方法得到意见领袖排名前 100 名用户的帖子,共有 160 个话题作为实验数据。在构造相关情感训练集时,利用本文 2.1 节中的方法进行构造,进而得到本文需要的训练集。分词采用清华大学开发的 CsegTag3.0 系统。

3.2 评估指标

对于意见领袖识别采用类似信息检索的评估方法来进行评价^[9]。采用的评估指标如下:

前 N 个结果中的正确率 $P@N$:

$$P@N = \frac{\text{前 } N \text{ 个结果中人工判定为真正意见领袖的个数}}{N}$$

前 N 个结果的平均正确率 $\text{Avg } P@N$:

$$\text{Avg } P@N = \left(\sum_{i=1}^N P@i \right) / N$$

实验中,两名工作人员根据用户所发全部帖子内容,以及用户的活跃度、受关注度和影响力覆盖度这三个统计指标来判断某一用户是否为一名真正的意见领袖。

3.3 实验目的及方法

实验目的是验证本文提出的考虑帖子倾向性的意见领袖识别方法的有效性。

实验方法是在使用相同的实验数据集和性能评估指标的前提下,将本文提出的考虑帖子倾向性的意见领袖识别方法与其他两种方法进行性能比较和结果分析。具体方法设计如下:

a) IDPM_NOSEN, 没有考虑帖子倾向性的 IDPM 方法。

b) IDPM_SEN_UNRE, 使用非相关训练集进行情感判别的考虑帖子倾向性的 IDPM 方法。

c) IDPM_SEN_REL, 使用相关训练集进行情感判别的考虑帖子倾向性的 IDPM 方法。

通过随 N 变化的 $P@N$ 曲线图来定性评估不同方法的效果,根据 N 取不同值时的 $\text{Avg } P@N$ 来定量评估不同方法的效果。

3.4 实验结果及分析

从图 2 可以看出,使用相关训练集判定帖子倾向性的 IDPM_SEN_REL 方法优于 IDPM_SEN_UNRE 和 IDPM_NOSEN 方法。

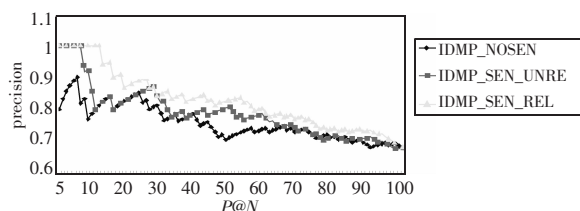


图2 三种方法 100 个结果的正确率变化曲线

表 1 是前 50 名用户中三种方法筛选出的真正意见领袖所占的比例。从表 1 可以看出,考虑帖子倾向性的 IDPM_SEN_UNRE 和 IDPM_SEN_REL 优于 IDPM_NOSEN 方法,这说明考虑帖子倾向性识别意见领袖是合理的; IDPM_SEN_REL 优于 IDPM_SEN_UNRE 和 IDPM_NOSEN 方法,其前 50 名识别正确率较 IDPM_NOSEN 和 IDPM_SEN_UNRE 方法分别提高了 13.89% 和 2.5%,这说明本文方法是有效的。

表1 前 50 名用户对比结果

方法	判断正确/个	相比 IDPM_NOSEN/%	相比 IDPM_SEN_UNRE/%
IDPM_NOSEN	36	-	-
IDPM_SEN_UNRE	40	+11.11	-
IDPM_SEN_REL	41	+13.89	+2.5

在本文方法的实验结果排名中,真正意见领袖的网络用户排名向前靠拢了,而非真正意见领袖的网络用户排名向后靠了。例如:用户“郑凯_杨苏剑”的排名由原来的第 5 名降到了第 17 名,而用户“郭晋嘉_郑欣”的排名由原来的第 10 名升到了第 6 名。观察语料可知,“郑凯_杨苏剑”的话题树中,在其所有的回复帖子中,负倾向性的回复帖子数占了总回复帖子数的 66.1%,这说明该楼主并非真正的意见领袖。而在“郭晋嘉_郑欣”的话题树中,其正倾向性的回复帖子数占总回复帖子数的 99.2%,说明该楼主是真正的意见领袖。从以上分析可知,考虑回复帖子的倾向性是有必要的,同时说明了本文方法的合理有效性。

表 2 ~ 4 分别是 N 取 100、50、20 时的平均正确率 $\text{Avg } P@N$,以便定量地评估 IDPM_NOSEN、IDPM_SEN_UNRE 和 IDPM_SEN_REL 这三种方法。

表2 Avg $P@100$ 的对比结果

方法	P 平均/%	相对 IDPM_NOSEN 的改变/%	相对 IDPM_SEN_UNRE 的改变/%
IDPM_NOSEN	76.65	-	-
IDPM_SEN_UNRE	79.65	+3.91	-
IDPM_SEN_REL	83.38	+8.78	+4.68

表3 Avg $P@50$ 的对比结果

方法	P 平均/%	相对 IDPM_NOSEN 的改变/%	相对 IDPM_SEN_UNRE 的改变/%
IDPM_NOSEN	81.73	-	-
IDPM_SEN_UNRE	86.01	+5.24	-
IDPM_SEN_REL	90.70	+10.97	+5.45

表4 Avg $P@20$ 的对比结果

方法	P 平均/%	相对 IDPM_NOSEN 的改变/%	相对 IDPM_SEN_UNRE 的改变/%
IDPM_NOSEN	86.36	-	-
IDPM_SEN_UNRE	93.04	+7.74	-
IDPM_SEN_REL	98.66	+14.25	+6.04

从表 2 ~ 4 可以看出,考虑了回复帖子倾向性的 IDPM 的 IDPM_SEN_REL 优于 IDPM_NOSEN 和 IDPM_SEN_UNRE 方法,其 Avg $P@50$ 相对于 IDPM_NOSEN 和 IDPM_SEN_UNRE 方法分别提高了 10.97% 和 5.45%。

(下转第 2636 页)

计算结果如图1所示。该图描绘了由 t_1 和 t_2 给出的Pareto前沿,Pareto解的数量为100,计算时间为5266.90 s。Pareto解的逼近性和分散性性能指标GD-metric和SS-metric^[15]随进化代数的变化如图2,3所示。由于优化软件Sedumi的优化效果很好,经过一次进化,得到的解就是真正Pareto前沿中的点,所以逼近性指标在生成初始解时就达到最优,而且在进化过程中始终保持最优。分散性指标随着进化代数的增加大体上有减小的趋势,说明进化后得到的Pareto解分散更为均匀。如果对分散性性能要求不高,在实际的计算中可以设置较小的进化代数就能得到很好的Pareto解集。

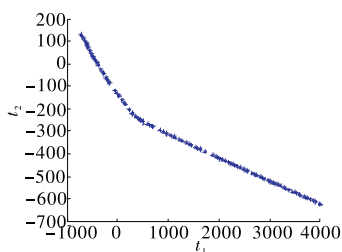


图1 计算得到的Pareto前沿

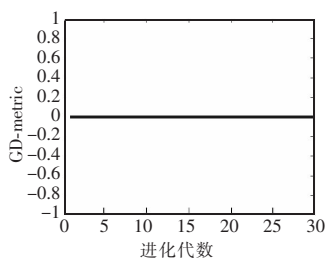


图2 GD-metric随进化代数变化

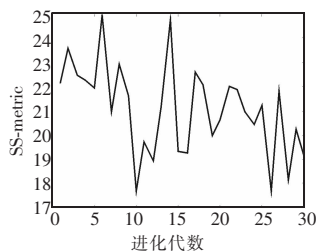


图3 SS-metric随进化代数变化

4 结束语

在实际决策过程中应用多目标规划模型时往往很难精确给出参数的确定数值,而目标函数一般来说对于优化模型的参数扰动十分敏感,决策者用不准确的参数计算规划模型时会面临较大的风险。因此,本文基于多个决策目标需要同时决策且均需获得鲁棒解的前提假设下,引入鲁棒多目标线性规划的定义,对椭圆扰动集下的不确定模型转换为具有二阶锥约束的确定性模型,并提出了基于智能优化和SOCP优化软件集成的

混合策略求解方法,通过数值算例验证了该算法的有效性。该算法的设计思路对于求解含有锥约束的多目标优化问题具有很好的借鉴作用。

参考文献:

- [1] COELLO C A C. Handling multiple objectives with particle swarm optimization [J]. *IEEE Trans on Evolutionary Computation*, 2004, 8(3): 256-279.
- [2] YAGMAHAN B, YENISEY M M. Ant colony optimization for multi-objective flow shop scheduling problem [J]. *Computers & Industrial Engineering*, 2008, 54(3): 411-420.
- [3] 丁力平, 谭建荣, 冯毅雄, 等. 基于Pareto蚁群算法的拆卸线平衡多目标优化 [J]. *计算机集成制造系统*, 2009, 15(7): 1406-1413.
- [4] JOZEFOWIEZ N, SEMET F, TALBI E G. An evolutionary algorithm for the vehicle routing problem with route balancing [J]. *European Journal of Operational Research*, 2009, 195(3): 761-769.
- [5] 谢涛, 陈火旺, 唐立山. 多目标优化的演化算法 [J]. *计算机学报*, 2003, 26(8): 997-1003.
- [6] 张勇德, 黄莎白. 多目标优化问题的蚁群算法研究 [J]. *控制与决策*, 2005, 20(2): 170-174.
- [7] SOYSTER A L. Convex programming with set-inclusive constraints and applications to inexact linear programming [J]. *Operations Research*, 1973, 21(5): 1154-1157.
- [8] BEN-TAL A, NEMIROVSKI A. Robust optimization—methodology and applications [J]. *Mathematical Programming*, 2002, 92(3): 453-480.
- [9] SUNGUR I, ORDONEZ F, DESSOUKY M. A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty [J]. *IIE Transactions*, 2008, 40(5): 509-523.
- [10] BURKE E K, De CAUSMAECKER P, De MAERE G, et al. A multi-objective approach for robust airline scheduling [J]. *Computers & Operations Research*, 2010, 37(5): 822-832.
- [11] LOBO M S, VANDENBERGHE L, BOYD S, et al. Applications of second-order cone programming [J]. *Linear Algebra and its Applications*, 1998, 284(1-3): 193-228.
- [12] 阎爱玲, 修乃华. 不确定信息多目标线性优化的鲁棒方法 [J]. *数学的实践与认识*, 2012, 42(6): 243-252.
- [13] 李砚, 杜纲. 椭圆不确定集下的鲁棒线性双层规划 [J]. *系统工程*, 2011, 29(11): 96-100.
- [14] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II [J]. *IEEE Trans on Evolutionary Computation*, 2002, 6(2): 182-197.
- [15] KNOWLES J, CORNE D. On metrics for comparing nondominated sets [C]//Proc of Congress on Evolutionary Computation. [S. l.]: IEEE Press, 2002: 711-716.

(上接第2615页)

4 结束语

随着新兴媒体逐渐地融入人们的日常生活,网络舆论影响力也在不断加大,有效地筛选出真正的网络意见领袖成为一项具有现实意义的任务。本文针对网络用户谈论的主题变化频繁、难以预先构造文本倾向性判定训练集的问题,提出了一种意见领袖识别中的倾向性判定方法,进而建立考虑帖子倾向性的意见领袖识别模型。该模型是建立在IDPM基础上的,并引进了考虑帖子倾向性的有效系数。相对于没有考虑帖子倾向性的IDPM_NOSEN方法和使用非相关训练集IDPM_SEN_UNRE方法,其前50名意见领袖发现的平均准确率分别提高了10.97%和5.45%,这说明了本文方法能够有效地提高意见领袖识别的正确率。

参考文献:

- [1] 陈然. 网络论坛舆论领袖筛选方法初探 [D]. 武汉: 华中科技大学

学, 2009.

- [2] 刘建明. 舆论传播 [M]. 北京: 清华大学出版社, 2001.
- [3] 樊兴华, 赵静, 方滨兴, 等. 影响力扩散概率模型及其用于意见领袖发现研究 [J]. *计算机学报*, 2013, 36(2): 360-367.
- [4] HUANG X J, ZHAO J. Sentiment analysis for Chinese text [J]. *Communications of CCF*, 2008, 4(2).
- [5] LIU H Y, ZHAO Y Y, QIN B, et al. Target extraction and sentiment classification [C]//Proc of the 10th Chinese National Conference on Computational Linguistics. 2009.
- [6] WILSON T, WIEBE J, HWA R. Recognizing strong and weak opinion clauses [J]. *Computational Intelligence*, 2006, 22(2): 73-99.
- [7] 樊兴华, 王鹏, 周鹏. 一种基于扩展的两步文本倾向性分析方法 [J]. *计算机工程与应用*, 2012, 48(1): 162-165, 169.
- [8] 樊兴华, 孙茂松. 一种高性能的两类中文文本分类方法 [J]. *计算机学报*, 2006, 29(1): 124-131.
- [9] FAN Xing-hua, NIE Jian-yun. Link distribution dependency model for document retrieval [J]. *Journal of Information & Computational Science*, 2009, 6(3): 1553-1564.