

文章编号: 1000-5862(2013)02-0136-06

# 基于群体智慧的电影排序模型

何文译, 林鸿飞\*, 杨 亮

(大连理工大学信息检索研究室 辽宁 大连 116024)

**摘要:** 利用新浪微博的微信群信息, 在蚁群算法的基础上, 提出了一种基于群体智慧思想的排序模型 (ACOR), 通过结合用户的偏好及电影热度对电影进行综合排名. 同时考虑了微博中的情感因素, 分析用户对电影的情感倾向并计算其情感值. 最后, 根据群体形成的情感积累值对热议的电影排序. 实验结果表明: 该模型更符合用户的偏好, 并具有一定的实时性, 可以有效地为用户提供相关电影信息.

**关键词:** 群体智慧; 情感因素; 微博; 蚁群算法; 电影排序

**中图分类号:** TP 391

**文献标志码:** A

## 0 引言

随着 Web2.0 尤其是以 Facebook、Twitter 为代表的社交网站的飞速发展, 越来越多的人在社交网络上发布和获取信息, 社交网络成为现实生活的网络延伸. 社交网络是指人与人之间通过血缘、兴趣、朋友等关系建立起来的社交网络结构, 根据兴趣、话题等凝聚形成的群体, 都是社交网络的范畴. 群体智慧也随着社交群体的形成越来越受人们的关注, 成为一个新热点. 分析社交群体的观点、行为等满足群体或用户的需求也愈加重要.

群体智慧应用的研究刚刚起步, Juho<sup>[1]</sup> 针对群体智慧的概念进行整合, 并总结为 3 类; M. Daly 等<sup>[2]</sup> 提出通过标签数量、用户信誉度等衡量文档信誉度, 对文档进行排名, 体现了群体智慧的思想; Pavlin Mavrodiev 等<sup>[3]</sup> 分析社交网络中群体智慧的影响力时, 对群体智慧的重要性给予肯定. 在观点挖掘和情感倾向性挖掘的研究领域中, Marreiros<sup>[4]</sup> 在分析群体决策时将情感因素细分为多类, 并提出情绪衰减以及情绪蔓延等对群体决策产生的影响. 中文方面吕韶华等<sup>[5]</sup> 提出基于 LDA 的评论倾向性排序. 社交网络具有实时性和群体性的特点, 本文将结合群体智慧针对微博进行研究.

本文结合群体智慧的思想, 将排序问题映射到

蚁群算法, 提出一种基于蚁群算法的排序模型 ACOR (Ant Colony Algorithm Rank Model), 该模型结合微博文本以及回复、转发等关系, 将微博的情感因素作为其中一部分, 分析群体对电影的态度变化以及大部分用户的兴趣偏好, 为更多用户提供相关信息.

## 1 理论基础

### 1.1 群体和群体智慧

本文借鉴徐旭林<sup>[6]</sup> 给出的“群体”定义: 在特定的时间内聚集在特定空间的一群物种, 如果其具有共同的动机和行为, 则称之为群体 (Crowd). 群体具有共同的动机和行为, 自发组成并且通过自组织的方式控制群体行为等, 个体之间有互相依赖和互相帮助等关系.

Juho<sup>[1]</sup> 指出“群体智慧”的一般定义是指通过个体之间的合作或竞争等呈现出的一种具有普遍形式和分散性的智慧, 也指学习由大量个体组成的群体的行为, 而这个群体中的个体又会受到其他个体和环境的影响.

### 1.2 蚁群算法

蚁群算法是群体智慧思想在计算机科学中的一个最为典型的实例和应用, 已在很多领域应用并取

收稿日期: 2012-11-15

基金项目: 国家自然科学基金 (61272370, 60973068) 和辽宁省自然科学基金 (201202031) 资助项目.

通信作者: 林鸿飞 (1962-), 男, 辽宁大连人, 教授, 博士生导师, 主要从事搜索引擎、文本挖掘、情感计算和自然语言理解等方面的研究.

得成功. 算法包括适应阶段和协作阶段. 适应阶段各候选解积累信息素, 协作阶段则通过信息素交流, 以期得到更好的解. 算法的主要思想: 蚁群中每个蚂蚁都可以释放并感知信息素, 在觅食的过程中会在经过的地方留下信息素, 信息素由于环境、时间等因素逐渐消失, 而其他的蚂蚁觅食的过程中则会倾向于信息素多的方向觅食, 如此循环, 在食物多的路径上会有积累更多的信息素, 从而找到最优的觅食路径等. 关键细节如下: (I) 环境: 包括障碍物、信息素、每个蚂蚁可感知环境信息、信息素以某速率蒸发. (II) 觅食: 蚂蚁感知范围内信息素多的地方觅食, 并不是向信息素最多的点移动. (III) 移动规则: 每只蚂蚁都朝信息素对多的方向移动, 没有指引则按自己原方向运动.

### 1.3 相关定义

对本文使用到的名词作如下定义.

用户信誉度: 用户的微博影响力和可信度, 由用户发微博数以及微博受关注程度决定.

情感值: 用户微博中流露出的情感, 由用户微博及用户信誉度确定, 其中微博的情感权重由情感倾向性和情感词的权重决定.

情感积累值: 针对某一对象(电影), 多个用户的微博情感值随时间的累加值, 每次衰减后在其衰减值上再次累加.

## 2 基于群体智慧的排序模型

### 2.1 ACOR 模型映射

本文中主要以蚁群算法为依托来利用群体智慧的思想做进一步研究. 一个自发形成的群体是实验的前提, 社交网络中因兴趣等组成的群体符合群体的特点. 社交群体的个体则相当于单个蚂蚁, 多个个体由特定枢纽构成一个群体类似于蚁群. 模型映射关系如图1.

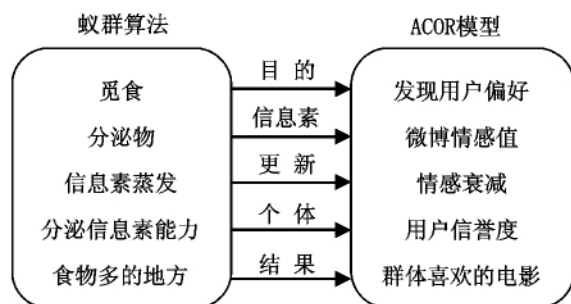


图1 模型映射

(I) 目的: 社交群体以发现共同的兴趣为目的, 本文选用的电影群, 目的为发现用户偏好.

(II) 个体: 每个个体信息的可信度是有区别的, 正如蚂蚁觅食中会存在分泌信息素的能力差异, 本文采用用户信誉度来衡量个体能力.

(III) 信息素: 用户微博中会流露出一定的情感, 本文以情感值作为信息素. 模型中则对一部电影微博数越多, 说明该方向越受关注, 情感值越大则电影越可能更受关注.

(IV) 更新: 心理学有情感衰减现象, 本文用情感值的衰减来模拟信息素蒸发. 由于微博的实时性和短周期性的特点, 用个体活跃程度衡量信誉度, 用户信誉度也随时间衰减.

(V) 结果: 群体的移动方向是情感值积累值最高的方向, 结果根据情感值积累值进行排名.

### 2.2 电影多元化

为了更准确把握人们对电影的情感, 本文中为每部电影建立一个多元组, 形式如下:

< 人物, 情节, 整体, 情感 >

其中人物指演员、导演等, 通过人物名识别; 情节包括剧情、效果等, 由标志词识别; 整体是对电影的整体评论, 情感则是微博中人们对该电影的情感. 电影做此多元化处理, 主要考虑到微博的简短性和无规则性, 很多微博只从单方面评价电影, 其情感值应有区别, 电影微博各方面的重要程度不同. 故本文由公式(1)计算每条微博的情感值.

$$E_{movie} = E_{role} \times \alpha + E_{detail} \times (1 - \alpha), \quad (1)$$

其中  $E_{movie}$  即电影情感值,  $E_{role}$ 、 $E_{detail}$  表示对电影人物和情节的情感值,  $\alpha$  即所占权重. 本文中  $\alpha$  取值 0.4, 即认为情节比人物相对更全面地评价电影. 如果人物和情节同时出现或都未出现则为整体评价, 不再考虑权重. 排序依据为所有微博的情感积累值.

### 2.3 模型具体介绍

2.3.1 信誉度的计算 M. Daly 等<sup>[2]</sup>提到用户信誉度主要由其他用户所决定, 如果某用户留下的信息别的用户关注的较多意味着该用户具有较高的信誉度, 同时, 用户留下的信息越多信誉度越高. 即留下信息越多且越受关注则用户的信誉度越高, 信誉度的计算公式(2)为

$$R_{new} = R_{old} + (1 - R_{old}) \times \gamma, \quad (2)$$

其中参数  $R$  表示信誉度, 因微博的实时性等特点, 回复数和转发数分布很不均匀, 且很多回复转发数均为 0, 为解决这一问题, 本文对参数做进一步规

范,计算公式如下为

$$\gamma = e^{\tau-1}, \quad (3)$$

其中  $\tau$  转发数确定,回复转发数越多值越大,且  $\tau \in [0, 1]$ . 由此  $\gamma \in [e^{-1}, 1]$  且  $\tau$  越大,  $\gamma$  越大. 故用户发微博时,其信誉度就会增大,同时其他用户关注越多,信誉度增大幅度就会越大. 语料回复、转发数统计结果如表 1,看出回复数小于 100,转发数小于 20 可以覆盖几乎整个语料,故本文中  $\tau$  值为微博实际回复、转发数分别与 100、20 的比值的均值.

表 1 回复、转发数统计

范围 \ 类型	= 0	1 ~ d	> d
回复数( d = 100)	51.64%	48.20%	0.16%
转发数( d = 20)	94.50%	5.48%	0.02%

2.3.2 信息素的计算 本文中的信息素由用户微博中流露出的情感来计算,针对微博内容字数较少,语言灵活的特点,使用由大连理工大学信息检索实验室的情感词汇本体<sup>[19]</sup>作为极性词典.情感词汇本体中情感词的极性分为褒义、贬义和无情感 3 类,共有 27 352 个词,其中包含褒义词 11 225 个,贬义词 10 855 个,无情感词 5 122 个.定义词典中词语的极性为

$$P_w = \begin{cases} 1, & w \text{ 为褒义,} \\ -1, & w \text{ 为贬义.} \end{cases} \quad (4)$$

情感词的情感强度不同,本文采用 TF-IDF 方法计算情感词的权重,以每个用户作为一个文档计算情感词的 tf-idf 值,同时权重值加入极性,即贬义则权重取负.

2.3.3 信息素的更新 信息素更新即情感值蒸发衰减的过程.在 Marreiros<sup>[3]</sup>等提到情感的衰减,情感本身是短暂且逐渐衰减的.情感衰减逐步趋于平静状态,符合信息素蒸发的特点.在心理学上情感的衰减曲线与指数函数曲线较为接近.在前人的基础上改动情感值随时间衰减公式为

$$E_t = E_{t-1} \times e^{-a \times t_i}, \quad (5)$$

其中  $E_t$  为  $t$  时刻的情感强度;参数  $a$  为情感衰减系数,且  $a \in [0, 1]$ ,主要用于控制衰减的速度,  $a$  值越大衰减速度越快;  $t_i$  为时间间隔.

2.3.4 信誉度的更新 用户信誉度的衰减来源于用户兴趣衰减,人们兴趣变化的研究中,郭进利<sup>[9]</sup>发现评论的时间间隔服从幂律分布,张红霞等<sup>[10]</sup>引入了艾宾浩斯的遗忘规律.本文对实时的微博用户信誉度的衰减引入艾宾浩斯遗忘曲线,如图 2,可以看到遗忘曲线比较符合幂律形式的变化规律,用户对事物的心理变化符合遗忘规律,因此在信誉度上

的变化也遵循这一规律.该曲线比较符合指数函数的特性,故本文中用户信誉度随时间的变化公式为

$$R_t = R_{t-1} \times e^{-b \times \Delta t}, \quad (6)$$

其中  $R_t$  为  $t$  时刻的信誉度;  $\Delta t$  为时间间隔;  $b$  为衰减系数且  $b \in [0, 1]$ ,  $b$  值越大衰减速度越快.可以看出公式(5)与公式(6)相似,因都符合指数变化,实际中衰减系数不同.

### 3 实验与分析

#### 3.1 实验语料

实验语料是新浪微博中“爱看电影”微群的微博信息,包含 4 月 15 日到 6 月 7 日微群中所有的微博,经去重、删除无用微博(如广告)等基本处理后,语料共包含 43 032 条微博,包括用户 ID、微博、更新时间、转发数和回复数.实验中抽取的电影是语料中评论数较多、较新的电影且是当时新浪风云榜(50 部)的交集,选出 10 部电影的微博数如表 2.在整个语料上训练情感词的 TF-IDF 权重,抽取包含表 2 电影的语料作为测试集.

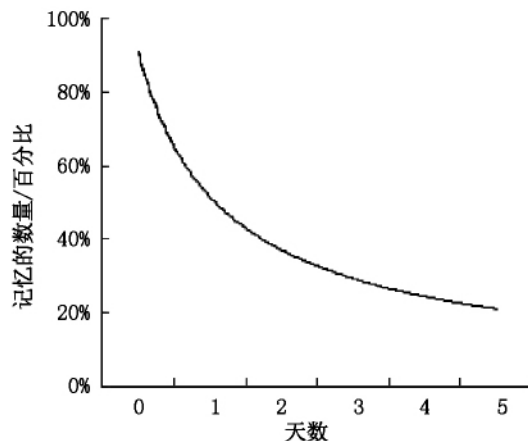


图 2 艾宾浩斯遗忘曲线

表 2 语料中各电影出现次数

电影	出现次数
复仇者联盟	2 787
春娇与志明	491
黄金大劫案	1 136
形影不离	221
普罗米修斯	102
晚秋	587
雨果	145
逆战	236
飞越老人院	371
超级战舰	1 701

本文实验预测排名与 6 月 6 日新浪风云榜的排名,以及采用新浪风云榜相同的方法(某时间段内

提及次数) 在语料统计得出的排名作对比, 以影评网站的评分排名作为标准排名. 本文排名加入了情感因素, 同时将历史数据作衰减后累加到情感累积值中排名.

3.2 评价指标

本文使用预测中衡量误差的平均绝对误差百分比 MAPE( mean absolute percentage error) ,公式( 7) 来衡量模型的预测准确率<sup>[11]</sup>

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|Pred_i - True_i|}{True_i}, \quad (7)$$

其中  $n$  表示排序的电影数量,  $Pred_i$  代表实验预测的排名,  $True_i$  代表实际排名. 从公式本身可以看出, 预测排名与实际排名越接近, MAPE 的值就越低, 也即预测的越准确.

3.3 实验结果及分析

为了说明 ACOR 模型具有一定的实时性, 选择

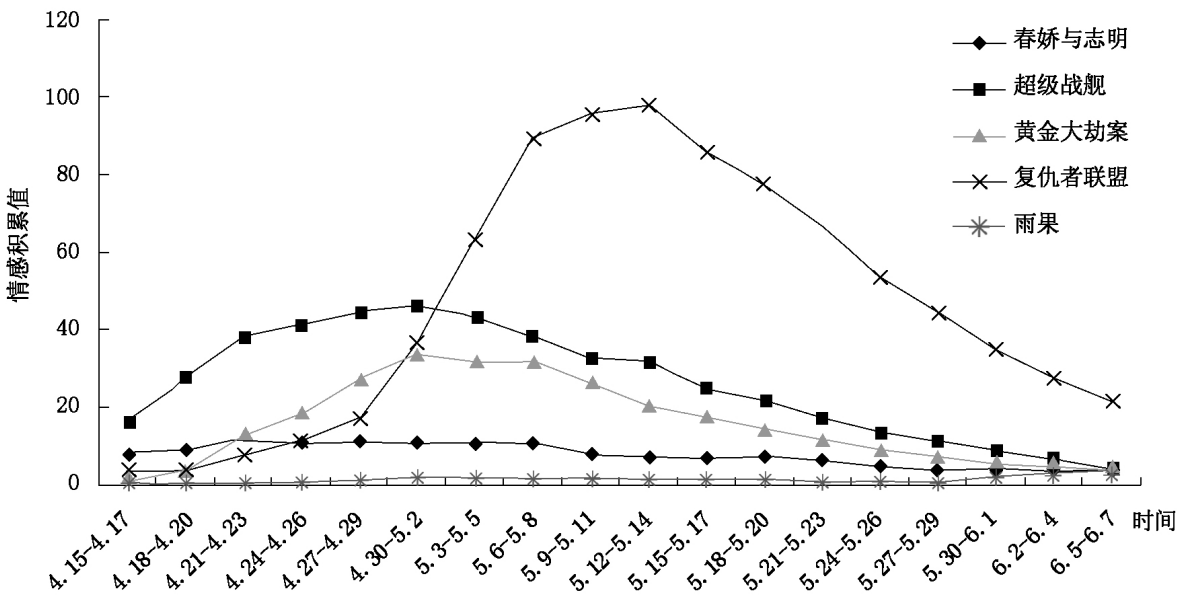


图3 电影实时情感累积值

图3 曲线反映出每部电影上映前后在每个时段的情感及热度变化, 不同的电影根据微博情感累积值显示出不同的热度. 如“复仇者联盟”明显比其他电影热. 电影上映前对电影的宣传等会带来一定的期待; 上映前后, 情感累积值有很明显变化; 而上映一段时间之后, 其情感累积值保持相对较平稳的状态. 如“春娇与志明”上映比语料时间早, 其值较稳定. 图3 说明该模型较符合微博实时性强更新快的特点, 具有一定的实时性.

本文以豆瓣电影的评分排名作为标准排名, 来衡量人们对电影的喜好. 豆瓣电影的评分人群以网

5 部电影分析语料其实时信息, 电影上映时间如表3. 其各时间段信息素值分布曲线如图3, 时间段为衰减周期, 为了保证每周期有一定数量的微博( 100 条以上), 选取衰减周期为3 天, 测试集后两周每天微博数如表4.

表3 5 部电影内地上映时间

电影	上映时间
春娇与志明	2012-3-30
超级战舰	2012-4-20
黄金大劫案	2012-4-23
复仇者联盟	2012-5-4
雨果	2012-5-31

表4 测试集微博数/天

时间段	微博数						
6-1—6-7	61	64	63	48	36	33	56
5-25—5-31	58	41	43	29	37	33	35

民为主, 通过评分的形式直接对电影进行评价, 是人们对电影的直观评价; 豆瓣参与评分人数较多, 比如“复仇者联盟”有118 485 人参与评分, 较能体现大众的观点, 较为客观; 此外, 针对包含国产电影的评价上, 很多网站也以豆瓣为主要参考标准.

截止本文语料下载日期( 2012. 6. 7) , 豆瓣电影中人们对部分新片的评分结果, 以及本文实验预测的排名结果、新浪风云榜6 月6 日的排名结果、以及在实验语料上采用新浪排名方法排名得到的结果如表5 所示. 本文方法与对比实验排名结果的 MAPE 对比如图4.

表5 豆瓣及各方法排名

排名	豆瓣电影	语料用新浪方法	新浪风云榜	实验预测
1	复仇者联盟	复仇者联盟	复仇者联盟	复仇者联盟
2	普罗米修斯	雨果	春娇与志明	普罗米修斯
3	飞越老人院	春娇与志明	黄金大劫案	雨果
4	春娇与志明	黄金大劫案	形影不离	黄金大劫案
5	雨果	普罗米修斯	普罗米修斯	春娇与志明
6	黄金大劫案	飞越老人院	晚秋	超级战舰
7	晚秋	晚秋	雨果	晚秋
8	超级战舰	形影不离	逆战	飞越老人院
9	逆战	超级战舰	飞越老人院	逆战
10	形影不离	逆战	超级战舰	形影不离

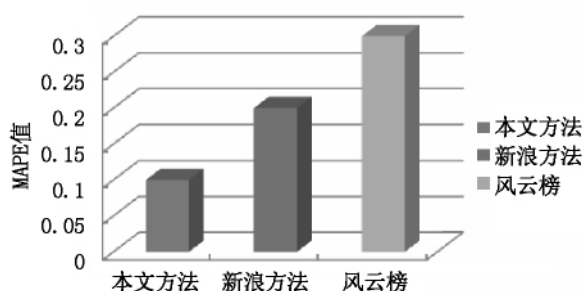


图4 MAPE 结果对比

图4表明本文ACOR模型的排名获得较低的排序损失值,比采用新浪微博方法排序的MAPE值降低了50%,比新浪风云榜排名降低更大幅度,主要原因在于Baseline方法只单纯考虑电影被提及的次数,而本文中加入了情感因素,考虑电影出现次数的同时也考虑微博中人们对电影的评价,由情感积累值表示信息素值,既体现电影被提及的次数,又体现人们对该电影的情感倾向,更准确的把握群体对电影的喜好程度。同时ACOR模型对微博的信息做了不同的处理,包括情感值累加、情感积累值随时间周期性衰减等,使得该模型在当前情感的同时将衰减之后的历史值也累加分析,使结果更符合大众的评价。此外,该模型可及时地反映人们对将要上映或正在上映的新片的态度以及期待程度,因微博的更新速度要比其他网站快很多,可以更早捕获人们的观点和偏好,提供更可靠的电影相关信息。本文的方法利用微博的历史信息,直接在衰减后的数值直接进行累加得到情感积累值,保留衰减后的数值下一周期计算时可直接累加,减少了实际应用中的复杂度。

本文实验得到的结果排名较高和较低的电影与豆瓣相符,而中间部分差别较大,“飞越老人院”差异最为明显。5月8日上映前情感积累值很低,上映后最高值仅为8.74,值一直不高,且语料中对该电影的评价多为感人或个人感悟,因电影本身就是对

生命的感悟、对老人的尊敬等。同时,模型中情感值的计算上也会可能带来一定误差。

## 5 结束语

本文针对新浪微博的微群信息在蚁群算法的基础上,提出一种基于群体智慧思想的排序模型(ACOR),并结合情感因素和心理学上的情感衰减现象,对用户的偏好及电影热度对电影排名。实验结果证明了ACOR模型有效性。本文中仅简单加入情感因素,下一步可以结合情感倾向分析中的方法对结果优化,同时根据电影上映前或上映中微博各时段的情感积累值,是否可以根据对电影票房、热度等信息做出预测,有待进一步研究。

## 6 参考文献

- [1] Juho Salminen. Collective Intelligence in Humans A Literature Review [EB/OL]. [2012-09-16]. <http://arxiv.org/abs/1204.3401>.
- [2] Elizabeth M. Daly harnessing wisdom of the crowds dynamics for time-dependent reputation and ranking [EB/OL]. [2012-09-12]. <http://www.computer.org/csdl/proceedings/asonam/2009/3689/00/3689a267-abs.html>.
- [3] Pavlin Mavrodiev, Claudio J Tessone, Frank Schweitzer. Effects of social influence on the wisdom of crowds [EB/OL]. [2012-09-13]. <http://arxiv.org/abs/1204.3463>.
- [4] Goreti Marreiros, Ricardo Santos, Carlos Ramos et al. Context aware emotional model for group decision making [J]. Intelligent Systems IEEE 2010(99):1-8.
- [5] 吕韶华, 杨亮, 林鸿飞. 基于LDA模型的餐馆评论排序[J]. 计算机工程 2011, 37(19):62-65.
- [6] 徐旭林. 社会群体行为建模及其动力学分析[D]. 天津:南开大学 2010.
- [7] Picard R. W. What does it mean for a computer to “have” emotions? [EB/OL]. [2012-09-15]. <http://hd.media.mit.edu/tech-reports/TR-534.pdf>.
- [8] 李海芳, 何海鹏, 陈俊杰. 性格、心情和情感的多层情感建模方法[J]. 计算机辅助设计与图形学学报 2011, 23(4):725-730.
- [9] 郭进利. 博客评论的人类行为动力学实证研究和建模[J]. 计算机应用研究 2011, 28(4):1422-1425.
- [10] 张红霞, 杨渊. 基于客户行为和兴趣变化的电子商务推荐系统[J]. 宝鸡文理学院学报:自然科学版 2012, 31

- (2): 1-5.
- [11] David A. Swanson, Jeff Tayman and T. M. Bryan. MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts [J]. Journal of Population Research 2011 28(2): 225-243.
- [12] David Martens, Bart Baesens, Tom Fawcett. Swarm intelligence for data mining [J]. Machine Learning (Machine Learning) 2011 82: 1-42.
- [13] 段海滨. 蚁群算法原理及其应用 [M]. 北京: 科学出版社 2005.
- [14] Jongwuk Lee, Seung-won Hwang, Zaiqing Nie, et al. Query result clustering for object-level search [C]. Paris: ACM, 2009: 1205-1214.
- [15] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造 [J]. 情报学报 2008 27(2): 180-185.
- [16] 大连理工大学信息检索实验室. <http://ir.dlut.edu.cn/>.

## A Movies Ranking Model Based on Collective Intelligence

HE Wen-yi, LIN Hong-fei\*, YANG Liang

(Information Retrieval Laboratory, Dalian University of Technology, Dalian Liaoning 116024, China)

**Abstract:** A ranking model based on the collective intelligence and which exploited the Ant Colony Algorithm to make the rank on the information of micro-group in the Sina Weibo has been proposed. At the same time, by combining with the extent of user's preference and the hot extent of movies, a comprehensive ranking is given. Besides, the emotion factor of micro-blogging is considered to analysis the users' emotion tendency and the value of the emotion is computed. The final ranking is based on the accumulative value of emotion about the hot movies in the group. The experiment shows that this model can meet the users' preference better than other method. And it has some degree of real-time properties to provide users with relevant information of movies effectively.

**Key words:** collective intelligence; emotional factor; micro-blogging; ant colony Algorithm; movie ranking

(责任编辑: 冉小晓)