掘客投票算法的属性论方法

许广林[1],刘念祖[1],冯嘉礼[2],刘永昌[2]
(1. 上海立信会计学院 数学与信息学院,上海 201620; 2. 上海海事大学 信息工程学院,上海 200135)

摘 要:掘客类型网站的技术核心是投票算法,而如何能够客观和公正地反应投票结果是投票算法的核心. 针对目前掘客类网站中投票算法过于简单,建立了一套投票算法指标体系,并且提出了基于属性论方法的投票算法,为掘客类网站以及其他类型的投票网站提供了一种新的投票算法. 给出实际例子的投票结果更公正和客观,从而有效地论证了算法的合理性.

关键词:投票算法;属性论方法;转换程度函数;评估模型
中图分类号: TP18 **文献标识码**: A **文章编号**: 1673-4785(2009)02-0118-04

A voting algorithm based on attribute theory for social news sites

XU Guang-lin[1], LIU Nian-zu[1], FENG Jia-li[2], LIU Yong-chang[2]
(1. College of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 201620, China; 2. College of Information Engineering, Shanghai Maritime University, Shanghai 200135, China)

Abstract: A key technology for a social news sites is the voting algorithm, which must objectively and fairly calculate voting results as its core requirement. In order to make the voting algorithms of social news sites more trustworthy, a new indexing system and new voting algorithm based on the methods of attribute theory was developed. An actual application involving such an algorithm was then developed, and it was discussed in the paper.

Keywords: voting algorithm; method of attribute theory; conversion degree functions; evaluating model

掘客是一种基于 Web 2.0的体现集体智慧[1-2]的新型网站,就像其他基于 Web 2.0的网站一样,推出不久就广受欢迎. 国外目前比较流行的掘客有 Digg、Reddit、Dzone、Dealigg和 Techtagg等,国内起步早的有来客掘客等. 掘客的核心功能就是提供一个发现和共享互联网资源和信息的平台,这个平台本身没有任何内容,网站所有的内容全部由用户提交,并且用户可以对其他用户提交的内容进行投票. 得票多少说明受用户喜欢的程度,假如得票数量达到一定程度,将会被置于首页显示,从而可以被网站所有的用户浏览. 由此可见,如何根据投票数对提交内容进行评分,也就是说,投票算法如何合理地反映内容的受喜欢程度,是掘客的技术核心所在.

通过调查和分析国内外几个掘客网站,发现它们在投票算法上都存在一定缺陷. 比如来客掘客、Dzone和 Dealigg,采用了类似半数投票算法[3],只是简单地对投票数进行累加,当某一内容的得票数累加到一定地步,就是被置于首页. 这种算法简单易用,但是针对复杂的互联网用户访问行为,存在 2个主要问题:第一,没有考虑投票用户经验值,也就是资深用户和新用户所投票是等价的,第二,没有考虑时间因素,比如来客常常有几个月以前提交的、已经过时的内容被置于首页. Digg考虑一些用户经验值和时间因素,但是对时间因素的考虑相对简单. 为了提高投票算法的合理性,本文在参考贝叶斯投票算法[4]和淘汰投票算法[5]基础上,应用属性论方法[6-7]的基本原理建立了一套新的投票算法.

收稿日期: 2008-12-16.
基金项目:上海市本级财政部门预算资助项目(1138IA0005).
通信作者:许广林. E-mail: glenxu@gmail.com.

1　评估指标建立

投票算法是一个复杂、多元的系统,要建立合理的投票算法,首先需要一个科学、合理的投票因素体系.根据掘客的基本特点,可以把投票因素分为 3个大部分共 5个要素.

1)内容提交者的经验值.包括内容提交者的忠诚度和内容提交者的活跃度 2个因素.忠诚度可以通过用户在线时间和用户的访问深度来衡量,它反映的是用户对网站内容的认可程度.活跃程度反映内容提交者的参与程度,它可以通过用户提交文章的数量和用户提交内容被置首页的比率来反映.

2)投票者的经验值.包括投票者的忠诚度和投票中的活跃度 2个因素.投票者的忠诚度体现了投票者对网站内容的认可和理解程度,可以通过用户在线时间和用户的访问深度来衡量.活跃度反映了投票者的参与度,可以通过投票中已投票数量的多少和被投票内容置首页的数量来反映.

3)投票数与时间的关系. Digg处理投票数和时间的关系,只是规定内容提交后 24小时,所有的投票失效,这种关系过于简单.而合理的做法是投票数和时间建立一个指数函数的关系,也就是说,2个得票数相同的内容,如果所获得票数的所用时间不同,那么它们的得分也应该不一样.

综合上述,评价指标体系如图 1所示.

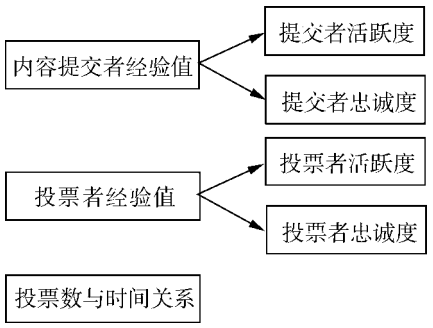


图 1　投票算法指标体系

Fig 1　Index system of vote algorithm

2　评估模型建立

投票系统的算法有如下步骤:

首先设 C为所有被提交的内容集合, sco_i为第 i个内容 c_i的综合投票分数, sco_i^{ce}为第 i个内容 c_i的提交者的经验值, sco_i^{ve}为第 i个内容 c_i的投票者的经验值, sco_i^{t}为第 i个内容时间趋势值.

1) 计算内容 C_i提交者的经验值.

设集合 X为网站所有 n个用户的集合,集合 Y为网站所有提交过内容的 m个用户的集合,集合 Z为网站所有投过票的 h个用户的集合,由此可得 $Z\subseteq X$和 $Y\subseteq X$.再设 t_i为某个用户在线时间, d_i为某个用户的访问深度,则根据转换程度函数[8],每个用户的忠诚度 l_i为

$$l_i(x)=\frac{1}{1+\exp\left(-\frac{n\cdot t_i}{\sum_{i}^{n} t_i}\right)}\cdot\frac{1}{1+\exp\left(-\frac{n\cdot d_i}{\sum_{i}^{n} d_i}\right)}. \tag{1}$$

设 $x_j\in Y,\ j\leq m$为第 j个提交过内容的用户,则 sc_j为此用户投票的数量, sr_j为用户提交内容被置首页的数量,则 x_j用户活跃度 sac_j为

$$sac_j=\frac{1}{1+\exp\left(-\frac{m\cdot sc_j}{\sum_{i=1}^{m} sc_i}\right)}\cdot\frac{1}{1+\exp\left(-\frac{m\cdot sr_j}{\sum_{i=1}^{m} sr_i}\right)}. \tag{2}$$

根据式 (1)和 (2),可以得出内容提交者的经验值 sco_i^{ce}为

$$sco_i^{ce}=(sac+l)/2 \tag{3}$$

2) 计算内容 c_i投票者的经验值 sco_i^{ve}.

设 $x_k\in Y,\ k\leq h$为第 k个投票的用户,则 vc_k为此用户提交文章的数量, vr_k为被用户投票的内容置首页的数量,则用户 x_k活跃度 vac_k为

$$vac_k=\frac{1}{1+\exp\left(-\frac{h\cdot vc_k}{\sum_{i=1}^{h} vc_i}\right)}\cdot\frac{1}{1+\exp\left(-\frac{h\cdot vr_k}{\sum_{i=1}^{h} vr_i}\right)}. \tag{4}$$

根据式 (1)和 (3),可以得出内容提交者的经验值 sco_i^{ve}为

$$sco_i^{ve}=\frac{vac+1\Big/\left(1+\exp\left(-\frac{k}{\sum_{i=1}^{k}(1-l_i)}\right)\right)}{2}. \tag{5}$$

被提交内容的投票数与时间的关系 $vt(t)$可以表示为

$$vt(t)=\begin{cases}\dfrac{1}{e^{t}}, & \text{if } t<1\,440\text{ m};\\ 0, & \text{if } t>1\,440\text{ m}.\end{cases} \tag{6}$$

3) 各投票因素权重的选择.

不同类型的掘客网站,针对的目标用户不同,要

求各个指标的权重也不相同. 比如 Digg网站,主要提供新闻,因此对时间因素考虑的更为重要,与此相反,Dzone主要以科学技术内容为主,因此对于投票者的经验值要求最高. 为了选择更加合理的权重,这里使用了属性论方法,它的基本思路是抛出几个样本点,由专家来进行评判,学习得出专家的心理曲线,从而得出各个指标的权重.

设 $score_0$ 为临界总分,在 $(score_0, 1)$中,根据曲线拟合要求,均匀选取若干个点: T_1、T_2、…、T_{n-1},在总分为 T_i $(i=1, 2, 3, \cdots, n-1)$的每个点上选取若干个样本让专家进行评分,按照式 (7)就可以找到总分为 T_i $(i=1, 2, 3, \cdots, n-1)$的重心坐标,而重心坐标反映了不同专家对各个投票因素的偏好.

$$b(\{c^h(z)\}) = \left(\frac{\sum_{h=1}^{t} V_1^h C_1^h}{\sum_{h=1}^{t} V_1^h}, \cdots, \frac{\sum_{h=1}^{t} V_m^h C_m^h}{\sum_{h=1}^{t} V_m^h} \right). \quad (7)$$

4)计算内容 c_i 的综合得分 sco_i.

$$sco_i = (\xi_1 sco_i^{ce} + \xi_2 sco_i^{ve} + \xi_3 sco_i^{t})/3. \quad (8)$$

5)设置一个阈值 sco_t,当内容 C_i 的综合得分大于或等于 sco_t,就把该内容置于首页.

3 实验分析

使用提出的投票算法,对来客掘客的投票算法进行了改进. 因为来客掘客更注重于新闻事件,所以首先给出样本数据,然后请几个新闻专家对样本数据进行打分,从而求出专家的心理重心曲线,最后得出各个指标所占的权重. 在使用此算法以后,关于新闻的内容虽然得票数和其他内容一样,但是它的得分更高,被置首页的概率更高. 图 2是使用此算法和没有使用此算法的被置首页内容比例分析图,从图 2可以看到,使用此算法以后,属于新闻事件的内容被置首页的比例从 22%提升到 36%.

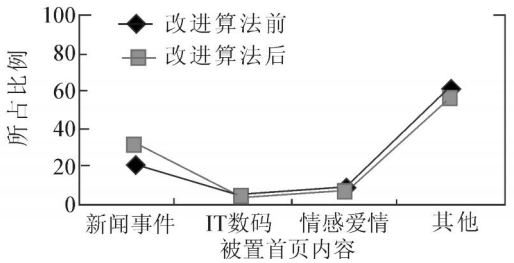


图 2 算法改进前与改进后被置首页比例

Fig 2 The results comparison table before and after algorithm improved

使用此算法后,得票数相同的内容,不一定得分相同. 如图 3中 (a)和 (b)内容都得 4票,并且都是一个用户所提交;但是因为 (a)的得分比 (b)高,所以 (a)被置首页,而 (b)没有. 究其原因,主要是内容 (a)在 30 min获得 4票,而内容 (b)是 240 min才获得 4票.

(a)使用来客掘客的投票算法

(b)未使用来客掘客的投票算法

图 3 被投票的内容

Fig 3 Voted contents

4 结束语

本文给出了一套投票算法的指标体系和基于属性论方法的投票算法. 该算法即能够体现网站的偏好,又能合理的反映投票者的经验值,为掘客类网站和其他带有投票功能的网站提供了一种新的方法. 同时给出的实际例子的投票结果也更加合理和公正,从而论证了本方法的合理性. 本文方法在考虑时间趋势值时使用的函数相对比较简单,下一步将尝试对多种势函数进行比较,从中择优.

参考文献:

[1] LÉVY. Collective intelligence: mankind's emerging world in cyberspace plenum [M]. New York: Plenum Trade, 1997: 37-39.

[2] SMITH J B. Collective intelligence in computer-based col-

laboration[M]. New York: Erlbaum, 1994: 132-135.

[3]BOYER R S, MOORE S. Mjrty—a fast majority vote algorithm[M]. The Netherlands: Kluwer Academic Publishers, 1991: 105-117.

[4]ALLEN C, APPELCLINE S. Collective choice: rating systems[EB/OL]. [2008-10-15]. http://www.lifewithalacrity.com/2005/12/collective_choi.html

[5]姚　昱, 朱山风. 基于投票模型的元搜索排序合成算法[J]. 计算机工程, 2007, 22: 214-216.
YAO Yu, ZHU Shanfeng. Voting model based on the sort of meta-search algorithm synthesis[J]. Computer Engineering, 2007, 22: 214-216.

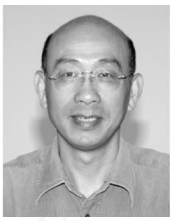
[6]FENG Jiali. Qualitative mapping orthogonal system induced by subdivision transformation of qualitative criterion and biomimetic pattern recognition[J]. Chinese Journal of Electronics, 2006, 15(6A): 850-856.

[7]FENG Jiali, MAO Qihuang, XU Guanglin, et al. Qualitative mapping, inner product transformation of qualitative criterion, artificial neuron and pattern[C]// Proc of Seventh IEEE International Symposium on Multimedia. Irvine, California, 2006: 15-20.

[8]冯嘉礼,许广林. 定性基准的线性变换与人工神经元网络[J]. 哈尔滨工程大学学报, 2006, 27(7): 6-12.
FENG Jiali, Xu Guanglin. Qualitative benchmark linear transform and artificial neural network[J]. Journal of Harbin Engineering University, 2006, 27(7): 6-12.

作者简介:

许广林,男, 1974年生,讲师,主要研究方向为机器学习、模式识别和风险评估. 发表学论文 10余篇.



刘念祖,男, 1955年生,教授,主要研究方向为数据库、数据挖掘.

冯嘉礼,男, 1948年生,教授、博士生导师. 中国人工智能学会理事、中国人工智能学会机器学习专业委员会副主任委员、中国管理科学研究院思维科学研究所学术顾问.

第 2届计算智能与设计国际学术研讨会
2009 International Symposium on Computational Intelligence and Design

On behalf of the successful symposium— ISCID 2008, the organizing committee and our local organizers wish to extend to you our personal welcome to attend the 2009 International Symposium on Computational Intelligence and Design (ISCID'09) which will be held at Changsha, China in 12～14, December 2009. It provides three day's focus on the science and technology that are the basis for the computational intelligence and design.

This symposium provides an idea-exchange and discussion platform for the world's engineers and academia, where internationally recognized researches and practitioners share cutting-edge information, address the hottest issue in computational intelligence and design, explore new technologies, exchange and build upon ideas. And provide researchers and practitioners interested in new information technologies an opportunity to highlight innovative research directions, novel applications, and a growing number of relationships between rough sets and such areas as computational intelligence, knowledge discovery and data mining, intelligent information systems, web mining, synthesis and analysis of complex objects, non-conventional models of computation and design.

We're certain you will find the city of Changsha and the surrounding area to be most pleasant and it will be my distinct pleasure to welcome each of you to the ISCID in December 2009.

Web site: http://www.iscid-conf.org/.