基于粗糙集-AHM的新浪微博意见领袖挖掘

□钮 亮 高 旭 雷园园
[中国计量学院 杭州 310018]

[摘 要] 传统上用来发掘意见领袖的方法主要有指标权重法和社会网络结构挖掘两大类，但是单纯靠指标权重法发现意见领袖受研究者的主观影响较大，而社会网络结构法中用户间关系较难挖掘且对用户其他属性的衡量有局限性，故而引入了基于粗糙集和AHM算法相结合的综合指标权重算法，充分综合主、客观指标权重挖掘意见领袖，避免了使用单一方法的弊端。通过对新浪微博中热点事件的实例验证，比较了粗糙集-AHM、AHP、社会网络挖掘三种算法的结果，并总结出了本方法计算简单，对用户关系数据依赖程度低、指标评价更加客观的特点。

[关键词] 意见领袖；粗糙集；AHM；指标评价
[中图分类号]G206 [文献标识码]A [DOI]10.14071/j.1008-8105(2016)01-0067-05

引言

微博是Web 3.0新兴起的一类开放互联网社交服务，它以集成化和开放化为特点，任何人都可以通过手机等多种途径向自己的微博客发布消息。微博以其发布内容的简明性、随意性、多样性和及时性为特点，领跑了真正的结构扁平、“去中心化”的自媒体时代。而随着微博时代的到来，意见领袖的也得到了越来越多人的重视。

意见领袖来源于Paul. Lazarsfeld的“两级传播”理论，是指在人际传播网络中可对他人施加影响的“活跃分子”，他们是信息传播的中介或过滤的环节，将信息传播给受众，形成信息传递的二级传播。随着意见领袖被越来越多的人所重视，国内外意见领袖的研究也在不断地发展。

常用的意见领袖发掘的方法主要可以分为指标打分法和社会网络结构挖掘两大类。指标打分法是指选取意见领袖的主要特征作为判定意见领袖的指标，如刘志明、刘鲁在研究新浪微博中意见领袖时以影响力、活跃度为一级指标，以被转发数、被评论数、原创数等7个特征为二级指标建立了指标体系[1]；丁汉青等人在发掘SNS中意见领袖时以中性、活跃度、吸聚力、传染性为四个一级指标和是否为管理员、好友数、关注数、被关注数、发帖数等12个为二级指标[2]。在指标体系建立后，可以通过层次分析法、评分函数模型等方法获得各指标的最终权重来发掘意见领袖。指标打分法包络面更广、可以根据侧重点不同有针对性的选取指标、操作也相对较为方便，但是指标及权重的确定在一定程度上受到研究者的主观影响。

社会网络结构挖掘是通过发掘用户间的相互关系来建立用户社会网络，根据用户在社区中的中心度及核心-边缘模型、影响力系数[3]来确定一个用户是否为意见领袖。Weng.J基于PageRank提出了Twitter-Rank方法以发现Twitter中有影响力的用户[4]；肖宇等人提出的LeaderRank算法在PageRank的基础上加入了情感权重[5]；薛可等人则借助于社会网络的相关理论研究了“意见领袖”在危机传播中的作用[6]。社会网络结构挖掘法在客观性方面有更大的优势，但是在新浪微博等网络社交平台中，用户间的相互关系较难挖掘，并且社会网络结构挖掘法对用户活跃度等的衡量有一定的局限。故而本文引入了一种新的算法用于意见领袖的挖掘即粗糙集-AHM[7]算法。

粗糙集-AHM算法是一种将粗糙集理论与AHM

[收稿日期] 2014－12－17
[基金项目] 浙江省高校人文社科重点研究基地基金(RWSKZD03-201207)；浙江省哲社重点研究基地和浙江省人文社科基金(SIPM3222)；浙江省社科联（2014Z084）；2014年度国家级大学生创新创业训练计划立项：基于社交网络的“杭州限牌”舆情分析模型构建与实证（201410356019）；2015年新苗人才计划项目（2015R409005）.
[作者简介] 钮亮（1975－ ）男，中国计量学院经济与管理学院教师；高旭（1994－ ）男，中国计量学院经济与管理学院本科生；雷园园（1993－ ）女，中国计量学院经济与管理学院本科生.

算法相结合的指标权重计算方法。粗糙集理论则作为一种能够定量化处理不精确、不一致、不完整信息的理论，最初由Pawlak教授与1982年提出[8]。经过三十多年的发展，而今的粗糙集理论被广泛地应用于机器学习[9]、数据挖掘[10]、决策支持[11]等众多方面。AHM算法是我国学者程乾生基于AHP提出的一套分析方法[12]。他在属性测度基础上，提出了相对属性测度和属性判断矩阵的概念，而相对权重和合成权重很容易从属性判断矩阵获得，故而AHM相对于AHP更为行之有效。将粗糙集-AHM算法引入意见领袖的挖掘，一方面能够充分综合主客观权重，避免了单单使用指标权重法主观影响较大的不足，使计算结果更为准确；另一方面粗糙集-AHM算法的计算结果可以对微博用户的基本属性例如活跃度、影响力等有直观的了解；更有效避免了社会网络结构法对数据要求较高的弊端。

一、相关理论

（一）粗糙集算法计算属性客观权重

粗糙集的简单定义[13]：设信息系统 $S=\{\mathrm{U},\mathrm{A},\mathrm{V},\mathrm{f}\}$，其中 U 是非空有限论域；A 为属性的有限集合，$A=C\bigcup D$，C 是条件属性集，D 是决策属性集；$V=U_{a\in A}V_a$，V_a 是属性 a 的值域；f：$U\times A\to V$ 是总函数，使得 $f(x_i,a)\in V_a$，对于每一个 $a\in A$，$\mathrm{x}_i\in U$。

设 $R\in A$ 是知识，且 $R\in Q$；$x_i,x_j\in U$，定义二元关系 $IND(r)$ 若满足以下条件：

$$IND(r)=\{(x_i,x_j)\in U\times U\mid \forall r\in R, r(x_i)=r(x_j)\}$$

则称 $IND(r)$ 是不可分辨关系，x_i 和 $\mathrm{x_j}$ 是关于属性R不可分辨的。

对于论域 U 上任意一个子集 X，X 不一定能用知识库中的知识来精确地描述，这时就用 X 关于 A 的上近似和下近似来“近似”的描述 X：

$$\underline{apr}\ \mathrm{X}=U\{[\mathrm{x}][\mathrm{x}]\ \subseteq \mathrm{X}\}=\{\mathrm{x}\in U\mid[\mathrm{x}]\subseteq \mathrm{X}\}$$

$$\overline{aprX}=U\{[x][x]\cap X\neq\phi\}=\{x\in U\mid[x]\cap X\neq\phi\}$$

；其中 $[x]$ 是 x 所在的 R 等价类。如果 $\underline{apr}\,\mathrm{X}\neq\overline{aprX}$，则称集合 X 是论域 U 上关于等价关系 R 的粗糙集。

$POSr(X)=aprX$ 称为 X 的 R 的正域；$NEGr(X)=U-aprX$ 称为 X 的 R 的负域。两个属性值 C 与 D 的依赖度 $q(C,D)$ 可以定义为：

$$q(C,D)=\frac{n(POSc(D))-n(POS\{C/ci\}(D))}{n(U)}\quad(1)$$

其中，$n(U)$ 表示集合 U 中元素的个数；$POS\{C/ci\}(D)$ 表示 D 的相对于 $\{C/ci\}$ 的正域。即将 U 的属性按照 $\{C/ci\}$ 进行分类以后，与 $POS(D)$ 的元素的交集。因此可以得知，$q(C,D)$ 的值越大，则其对应的指标权重越大；反之，则越小。我们便利用粗糙集的这一性质来计算属性指标的客观权重。

（二）AHM计算属性主观权重

AHM是一种无结构的多准则决策方法，将定性分析和定量分析相结合，将人们原本不系统的思维过程层次化、数量化，从对待解决问题的不同影响角度出发，将问题的影响因素一一列出并找出相应的隶属关系进而进行层次聚合，形成属性层次结构模型。AHM中的比较测度矩阵需要由AHP判断矩阵转化而来[14]。

利用AHM进行分析时需要先对待解决的问题构造阶递层次模型，包括目标层、准则层和方案层三个层级。设次准则层也即方案层有n个元素，分别将其记为 $A_1,A_2,A_3,\ldots A_n$。分别对准则层的准则 $C_1,C_2,C_3,\ldots C_n$ 比较两个不同元素 A_i 和 A_j $(i\neq j)$ 的相对重要性，分别记做 L_{ij} 和 L_{ji}。由属性测度，L_{ij} 和 L_{ji} 应满足如下要求：

$$L_{ij}\geq 0, L_{ji}\geq 0, L_{ij}+L_{ji}=1$$

由于元素 A_i 无法和自身比较相对重要性，故而规定 $L_{ii}=0$，$1\leq i\leq n$

$$\mathrm{wc}(\mathrm{i})=\frac{2}{n(n-1)}\sum_{j=1}^{n}Lij\quad \mathrm{i}=1,2,\ldots,\mathrm{n}\qquad(2)$$

$\mathrm{wc}=(w_{c(1)},w_{c(2)},\ldots\ldots\mathrm{wc}(n))^T$ 其中 W_c 为相对属性权向量。L_{ij} 的计算可由如下公式获得：

$$L_{\mathrm{ij}}=\begin{cases}\dfrac{\mathrm{k}}{k+1}, k=a_{ij}\\ 0.5\ ,\ i=j\\ \dfrac{1}{k+1}, k=\dfrac{1}{a_{ij}}\end{cases}\qquad(3)$$

其中k为大于2的正整数。元素 L_{ij} 和 L_{ji} 对准则 C 的比较可由层次分析法AHP中的相对比例标度 $\mathrm{a_{ij}}$ 给出，在准则 C 下，利用9标度法度量 A_i 和 A_j 的相对重要程度。

方案层中各因素与系统目标的合成权重，可由如下公式计算得到：

$$W=(W_{c1},W_{c2},\ldots,W_{cn})W_G\qquad(4)$$

（三）综合权重的计算

本文利用粗糙集-AHM算法分别计算出了各属性的主、客观权重，为了进一步获得更为合理和科学的指标权重，使挖掘出的意见领袖更为准确，本文采用综合的权重计算函数来对两组权重进行计算

得出最终权重。通过研究分析，客观的数据准确性较高，更遵从实际，并引进黄金分割定律[7]来构建综合权重计算函数，综合两种权重，得到最终评价指标权重：

$$W = \partial W_{ai} + (1-\partial)W_{bi} \quad (5)$$

其中，W_{ai}表示客观属性权重，W_{bi}表示主观属性权重。本文计算意见领袖的属性权重更偏向于客观事实依据，因此把黄金分割点的近似值0.68赋给∂，最后计算得到的W即为综合权重。

二、“马航失联”实例验证

（一）粗糙集计算客观权重

本文通过Rosetta软件来计算马航失联事件中用户的属性重要程度的权重。我们爬取了马航失联事件中共10475个用户的信息，作为验证的样本值，即样本集合$U=\{X_1, X_2, X_3, \ldots, X_{10474}, X_{10475}\}$，选取他们的粉丝数，关注数，转发、评论、认证、发博6个属性评价指标作为条件属性；是否为意见领袖作为决策属性；即：$C=\{C_1, C_2, C_3, C_4, C_5, C_6\}$；$D=\{$是意见领袖,不是意见领袖$\}$。

经过对于用户条件属性数据的分析，粉丝、关注、转发、评论等数据都是连续性数据，服从幂律分布，为了防止因样本数量过多而出现多数条件属性为0的情况，为了利用rosetta处理数据，我们要对数据进行离散化。以10为底对变量取对数，再对所得数据进行离散。粉丝数经过取对数离散后的结果([0,1],1),([1,2],2),([2,3],3),([3,4],4),([4,5],5),([5,6],6),([6,7],7),([7,8],8)；关注取对数离散的结果([0,1],1),([1,2],2),([2,3],3),([3,4],4)；评论取对数离散的结果是([0,1],1), ([1,2],2),([2,3],3),([3,4],4)；发博离散结果是([0,50],1),([50,100],2),([100,150],3),([150,200],4),([200, 250),5)。将离散以后的数据建立决策判断矩阵，如表1所示.

表1 决策判断矩阵表

序号	粉丝	关注	转发	评论	发博	认证	性别
1	7	2	1	1	1	是	男
2	7	2	1	1	1	是	男
…	…	…	…	…	…	…	…
N	4	2	3	2	2	否	女

1. 计算属性的等价类。利用粗糙集分别计算条件属性和决策属性的等价类通过Rosetta软件的Other-Partition功能来实现，结果如下：

$U/IND(C)$ ={{1}{2,91,106,109,123,165}{4784},…,{4788,10423,10475}}

$U/IND(D)$ ={{1,31,32,33,44,45,46,47,48,…,88,89},{23,24,25,26,27,…,362,363}}

接下来计算条件属性分别移去一个属性得到的等价类$U/IND(C-C_i)$：

$U/IND(C-C_1)$ ={{1},{2,29,47,52,61,62,63,66,91,106,109,123,165,238,248,249,255,256,746},…,{10436,10437,10438,…,10475}}

$U/IND(C-C_2)$ ={{1},{2,91,106,109,123,134,165},{3},…,{10463,10464,10465,…,10473,10474,10475}}

$U/IND(C-C_3)$={1},{2,73,91,106,109,123,126,165},…,{10453,10454,10455,10456,…,10474,10475}}

$U/IND(C-C_4)$ ={{1},{2,5,76,82,91,98,101,106,108,109,123,139,140,141,165},…,{10453,10454,…,10474,10475}}

$U/IND(C-C_5)$ ={{1},{2,91,106,109,123,165},{3},{4,120},…,{ 10453,10454, ,…, 10475}}

$U/IND(C-C_6)$ ={{1,7,79,81,84,85,87,92,93,94,95,96,99,100,112,114,121,127,132,138,142,143,146,150,153,155,156}, ,…,{9573,9903,9435,9600,9618,9630,9951,9952,…,10474,10475}}

2. 计算各个属性的正域

通过Rosetta软件Other-Approximate decision class功能来实现

$POSc(D)$={{1, 31, 32, 33, 44, 45, 46, 47, 48, 49, 50, …,87, 88, 89} }

$POS_{\{C/c1\}}(D)$={{1, 32, 33, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89}，{44}，{31, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72}}

$POS_{\{C/c2\}}(D)$={{58, 78, 88, 89}，{1, 33, 47, 48, 49, 50, 51, 52, 57, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 73, 76, 77, 79, 80, 81, 82, 83, 84, 85, 87}，{31, 32, 45, 46, 53, 54, 55, 56, 70, 71, 72, 74, 75, 86}，{44}}

$POS_{\{C/c3\}}(D)$={{1, 53, 65, 67, 68, 78, 79, 80, 81, 84, 85, 87}，{88, 89}，{31, 33, 47, 52, 54, 55, 56, 61, 62, 63, 64, 66, 69, 70, 71, 72, 76, 77, 82, 83}，{32, 44, 45, 46, 48, 49, 50, 51, 57, 58, 59, 60, 73, 75, 86}，{74}}

$POS_{\{C/c4\}}(D)$={ {53, 80, 88, 89}，{1, 31, 33, 54, 55, 56, 64, 65, 67, 68, 69, 70, 76, 77, 78, 79, 81, 82, 84, 85, 86, 87}，{32, 44, 46, 47, 48, 49, 50, 52, 57, 60, 61, 62, 63, 66, 72, 73, 75, 83}，{45, 51, 58, 59, 71, 74}}

$POS_{\{C/c5\}}(D)$={ {1, 31, 32, 33, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89}}

$POS_{\{C/c6\}}(D)$={ {33, 55, 56, 59, 77, 83, 88, 89}, {31, 32, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 78, 79, 80, 81, 82, 84, 85, 86, 87}, {1}}

3. 计算属性的重要性程度

按照公式（1）来计算属性的重要性程度，并做归一化处理得到各个属性的客观权重，结果如下：

表2 属性客观权重表

属性	C1	C2	C3	C4	C5	C6
权重	0.250	0.157	0.159	0.146	0.109	0.179

（二）AHM算法计算属性的主观权重

为了充分听取专家意见，设计用户属性评价指标体系问卷，分别设立了2个一级指标和6个2级指标，一级指标的设定参照现有的研究分为用户影响力和用户活跃度两个指标[3]，二级指标则选定了以上6个属性指标项，采用9标度法，由专家对两两属性的比重进行打分。利用专家评分来分别计算属性的主观权重，并求取平均值作为属性的主观权重。根据以上两级指标构建属性层次结构矩阵，如下图所示。

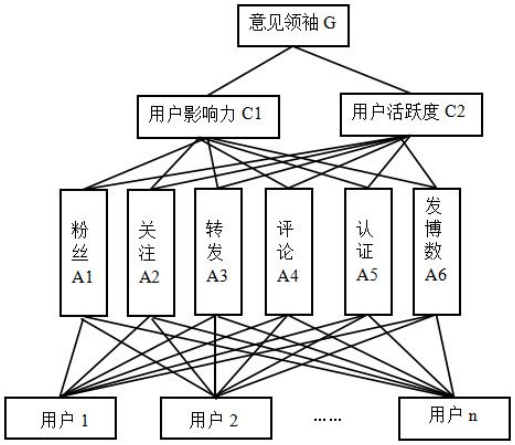

图1 意见领袖层次模型

1. 确定一级指标权重

通过专家的评分构建准则层结构矩阵来计算两个一级指标权重；

2. 根据公式（3）计算所有决策层二级指标分别对用户影响力和用户活跃度两个一级指标影响权重；

3. 根据公式（5）计算得出所有二级指标对于目标层的影响权重如下表：

表3 二级指标权重表

二级指标	粉丝	关注	转发	评论	认证	发博
权重	0.237	0.128	0.135	0.133	0.171	0.196

（三）确定综合权重

按照公式（6）计算最终评价指标综合权重：

表4 评价指标综合权重表

序号	C1	C2	C3	C4	C5	C6
指标	粉丝	关注	转发	评论	认证	发博
综合权重	0.248	0.160	0.158	0.146	0.110	0.178

粉丝的权重最高，因此粉丝对于能否成为意见领袖具有决定性作用，普通用户要想提升自己的影响力、成为意见领袖，首先应增加自己的粉丝量。

（四）计算意见领袖值

利用计算得到的属性指标综合权重来计算马航失联事件中用户的意见领袖值，为了精确计算，本文把认证属性进行量化，算法如公式（7）所示。

$$OL=u1\times W_{c1}+u2\times W_{c2}+\ldots+u6\times W_{c6} \quad (7)$$

其中$u1$到$u6$分别代表用户6个属性的属性值，最终计算出马航失联用户的意见领袖值，并取前1‰作为意见领袖，得到马航失联用户的意见领袖。

为了验证粗糙集-AHM算法在意见领袖寻找中的科学性和可靠性，本文同时利用AHP算法[3]和社会网络分析法[5]分别对马航意见领袖进行挖掘，并对3种方法的得到的意见领袖进行对比，如表5所示。

表5 利用3种方法得到的“马航”事件意见领袖

方法排名	粗糙集-AHM	AHP	社会网络分析法
1	姚晨	姚晨	姚晨
2	延参法师	延参法师	延参法师
3	人民日报	人民日报	思想聚焦
4	央视新闻	陈里	陈里
5	陈里	央视新闻	人民网
6	鞍钢郭明义	鞍钢郭明义	张颐武
7	人民网	人民网	鞍钢郭明义
8	潘石屹	黑人建州	潘石屹
9	黑人建州	潘石屹	微天下
10	微天下	微天下	袁裕来律师

从表中可以看出，利用粗糙集-AHM算法和AHP算法挖掘得到的马航意见领袖基本相同，而与利用社会网络分析法得出的意见领袖相比也有较高的吻合度，因而可以确定利用粗糙集-AHM算法来挖掘意见领袖是切实可行的。

三、总结

本文引用粗糙集-AHM算法来减弱由单纯AHM算法带来的主观性。用粗糙集算法对各指标求取客观权重，用AHM算法求取主观权重，再由所得的客

观、主观权重求取指标的综合权重值。利用综合权重来挖掘意见领袖。

本方法特点体现在两方面，首先，在指标评价法的基础上，通过引进粗糙集来代替以往的统计方法或专家经验，去除一定的主观因素，在对指标量化方面，AHM相较于AHP来说，有计算量小、模型简单、无需进行一致性检验、决策效率高等特点；其次，一定程度上降低了对数据的要求，在微博用户间关系的数据较难爬取的背景下提供了更为切实可行的意见领袖挖掘方法。

本方法是对指标评价体系的一种改进，对于通过指标评价法来挖掘领袖的案例均具有普适性，而且不仅局限于新浪微博，还可移植到其他社交平台。但由于粗糙集-AHM是在指标评价体系基础上建立起来的，对于用户的属性信息具有一定的依赖性；其次，每次都要重新计算用户属性权重，具有一定的局限性。

参考文献

[1] 刘志明, 刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程, 2011(06):8-16.

[2] 丁汉青, 王亚萍. SNS网络空间中“意见领袖”特征之分析——以豆瓣网为例[J]. 新闻与传播研究, 2010(03): 83-90,111.

[3] 李卓卓, 丁子涵. 基于社会网络分析的网络舆论意见领袖——以大学生就业舆情为例 [J]. 情报杂志, 2011(11): 66-70.

[4] WENG J S, LIN E P, JIANG J, et al. Twitterank: finding topic-sensitive influential twitterres[A] //Proceeding of the Third ACM International Conference on Web Search and Data Mining [C]. New York :2010, 261-270.

[5] 肖宇, 许炜, 夏霖. 一种基于情感倾向分析的网络团体意见领袖识别算法[J]. 计算机科学, 2012, 39(2): 34-37.

[6] 薛可, 陈晞, 王韧. 基于社会网络的品牌危机传播“意见领袖”研究 [J]. 新闻界, 2009(08): 30-32.

[7] 张晓明. 基于粗糙集-AHM的装备制造业企业创新能力评价指标权重计算研究[J]. 中国软科学, 2014(6): 151-158.

[8] PAWLAK Z. Rough sets[J]. .International Journal of Computer and Information Sciences, 1982, 11:341-356.

[9] CHMIELEWSKI M R, GRZYMALA-BUSSE J W. Global discretization of continuous attributes as preprocessing for machine learning[J].International Journal of Approximate Reasoning,1996, 15: 319-331.

[10] CHAN C C. A rough set approach to attribute generalization in data mining[J]. Journal of Information Sciences, 1998, 107: 169-176.

[11] PAWLAK Z. Rough set approach to Know ledge-based decision support[J]. European Journal of Operational Research, 1997, 99: 48-57.

[12] 程乾生. 层次分析法AHP和属性层次分析模型AHM [J]. 系统工程理论与实践, 1997(11): 56-59.

[13] 张文修, 吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学, 2000, 14(4): 1-12.

[14] 宿程远, 吕森, 赵旭雍,等. 属性层次分析模型在小城镇污水处理厂规划中的应用[J]. 广东农业科学, 2011, 38(2): 163-165.

Recognition of Opinion Leaders in Microblog Based on Rough Set and AHM

NIU Liang GAO Xu LEI Yuan-yuan
(China Jiliang University Hangzhou 310018 China)

Abstract Traditionally, there are two general types of methods to seek for an opinion leader, which are index weighting method and digging of social network structure. But simply relying on index weighting method to find an opinion leader would be affected by the subjective idea of the opinion leaders, while the method of digging out social network structure is also limited due to the seeming invisible relationship among the users and the unreliable evaluation of the other quality of the users. The article brings up an integrated index weighting algorithm based on rough set and AHM algorithm judging from the goods and weakness of the methods mentioned above. In this way, it is possible to have an integrated index of both subjective and objective standard to judging on the potential opinion leaders in the meantime avoiding the disadvantages of using one single method. The article has some exemplification research on certain hot hits on Weibo. By comparing the output of the three methods, it is concluded that this methods is featured as simple and objective.

Key words opinion leader; rough set; AHM; evaluation index

编辑 何 婧