DOI: 10.1051/jnwpu/20183620368

基于 Vague 软集相似度量的网络舆情综合评判方法

王伟[1], 武君胜[1], 朱志祥[2], 杨文超[3]

(1.西北工业大学 软件与微电子学院, 陕西 西安 710072;
2.西安邮电大学 物联网与两化融合研究院, 陕西 西安 710061;
3.西北工业大学 计算机学院, 陕西 西安 710072)

摘 要:针对现有 Vague 软集相似度量方法的局限性,修正了 Vague 软集相似度量公式,提出了一种考虑 Vague 值区间中心差异性的 Vague 软集相似度量方法,并给出公理化证明。基于大规模网络舆情数据集的综合评判分析实验表明,该方法是合理的、有效的、可行的,在网络舆情分析等综合决策问题研究中有较好的应用前景和效果。

关 键 词:Vague 软集;相似度量;网络舆情
中图分类号:TP18 **文献标志码**:A **文章编号**:1000-2758(2018)02-0368-07

Gau 和 Buehrer 在 1993 年提出的 Vague 集理论[1]是对 Fuzzy 集的补充和扩展。在处理不确定性信息时,Vague 集比传统的模糊集有更强的表达能力和灵活性,是一种新型的处理模糊性问题的数学分析模型。软集理论[2]是 Moldtsov 在 1999 年提出的一种新的处理不确定性和不精确性信息的数学工具,该理论引入了参数化思想,克服了 Vague 集只能处理部分不确定性信息的不足,在模式识别、数据挖掘、模糊决策、图像检索等实际问题中,有很大的应用潜力。上述 2 种理论都从不同角度聚焦信息系统中知识的不确定、不完备和不精准等问题,在实际应用时既相互联系又相互补充,因此可以进行融合,以发挥各自的优势,弥补各自的不足。针对 Vague 集和软集的融合问题,文献[3-6]将 Vague 集与软集理论进行结合,提出了新的 Vague 软集模型,并研究了相关性质及系列问题,目前已成为一个新兴的研究方向。在基于 Vague 软集的不确定信息处理中,判定 2 个 Vague 软集模式的相似度,是研究基于 Vague 软集的知识划分、模糊决策及综合评判等问题[7-9]的前提,吸引了众多研究者的关注。

分析发现,Vague 软集的本质是具有 Vague 集区间特征的软集。一个区间的特征,一般有 4 个重要的参数,即其左(右)端点、区间长度以及中点等。因此,在研究 Vague 软集的相似度量方法时应充分考虑 Vague 集的所有数值区间特征,包括真隶属度、假隶属度、犹豫度、核以及 Vague 值的区间中心等主要特征。现有文献给出的 Vague 软集相似度量公式,大多是从部分因素来衡量 Vague 软集的相似度量。如文献[10-11]提出的 Vague 软集相似度量衡量方法只考虑了 Vague 集的真隶属度、假隶属度以及核的差异性,却忽略了犹豫度和 Vague 值的区间中心等特征因素;文献[12]提出的 Vague 软集相似度量公式,重点考虑了真隶属度、假隶属度以及犹豫度的差异性,没有充分考虑 Vague 集核以及 Vague 值区间中心 2 个特征因素;文献[13]基于欧式距离提出了一种考虑真隶属度、假隶属度以及犹豫度差异性的 Vague 软集相似度量方法,忽略了 Vague 集核及区间中心 2 个特征因素;文献[14]引入参数权重提出一种 Vague 软集相似度量方法,但只考虑了 Vague 集真隶属度、假隶属度以及犹豫度的差异性;文献[15]提出的 Vague 软集相似度量公式只简单考虑了 Vague 集真假隶属度的差异性。本文在上述

收稿日期:2017-04-10 **基金项目**:博士后基金(2014M552478)与陕西省教育厅专项科研计划基金(14JK1724)资助
作者简介:王伟(1979—),西北工业大学博士后,主要从事人工智能、数据挖掘等方面的研究。

研究的基础上,将 Vague 值的区间中心这一重要特征引入 Vague 软集相似度理论进行研究,并给出了新的 Vague 软集相似度量的定义及公理化证明,同时将结果应用到网络舆情综合决策分析问题,对与此相关的一些关键问题进行了探索性研究,本文的相关研究结果,可为网络舆情评判等其他综合决策问题提供了理论基础。

1　预备知识

下面对有关基础理论进行描述。

1.1　Vague 软集

Vague 软集模型描述如下:

定义 1(Vague 软集)　设 U 是一个论域,E 是一个参数集,$A \subseteq E$,且 $F:A \to P(U)$ 是一个映射,即对 $\forall e \in A$,$F(e)$ 为 U 上的一个 Vague 集,称 (F,A) 为 U 上的一个 Vague 软集。

定义 2(Vague 软相等)　设 (F,A)、(G,B) 为 U 上的 2 个 Vague 软集,若 $A \subseteq B$,且对于 $\forall e \in A$,$x \in U$,有 $t_{F(e)}(x) \leqslant t_{G(e)}(x)$,$f_{F(e)}(x) \geqslant f_{G(e)}(x)$,则称 (F,A) 软包含于 (G,B)(或称 (G,B) 软包含 (F,A)),记作 $(F,A) \tilde{\subseteq} (G,B)$(或 $(G,B) \tilde{\supseteq} (F,A)$);若有 $(F,A) \tilde{\subseteq} (G,B)$ 且 $(G,B) \tilde{\supseteq} (F,A)$,则称 (F,A) 与 (G,B) Vague 软相等。

定义 3(Vague 软集的补集)　设 (F,A) 为 U 上的一个 Vague 软集,称 $(F,A)^c = (F^c, \neg A)$ 为 (F,A) 的补,其中 $F^c: \neg A \to V(U)$,即对于 $\forall \neg e \in \neg A$,$x \in U$,有:

$$t_{F^c(\neg e)}(x) = f_{F(e)}(x), 1 - f_{F^c(\neg e)}(x) = 1 - t_{F(e)}(x)$$

定义 4(相对空的 Vague 软集)　设 U 是一个论域,E 是一个参数集,$A \subseteq E$,(F,A) 为 U 上的一个 Vague 软集,若对 $\forall e \in A$,有 $x \in U$,$t_{F(e)}(x) = 0, 1 - f_{F(e)}(x) = 0$,则称 (F,A) 为 U 上的一个相对空的(相对于参数集 A)Vague 软集,记为 ϕ_A。

定义 5(相对全的 Vague 软集)　设 U 是一个论域,E 是一个参数集,$A \subseteq E$,(F,A) 为 U 上的一个 Vague 软集,若对 $\forall e \in A$,$x \in U$,有 $t_{F(e)}(x) = 1, 1 - f_{F(e)}(x) = 1$,则称 (F,A) 为 U 上的一个相对全的(相对于参数集 A)Vague 软集,记为 μ_A。

1.2　Vague 软集间的相似度量定义

文献[10]提出了 Vague 软集间的相似度量应满足的公理化定义:

定义 6　设 $VSS(U)$ 表示论域 U 上的 Vague 软集,E 是一个参数集,(F,E),$(G,E) \in VSS(U)$,函数 $M: VSS(U) \times VSS(U) \to [0,1]$ 称为 Vague 软集间的相似度量。如果其满足以下条件:

准则 1　有界性:$M((F,E), (G,E)) \in [0,1]$;

准则 2　对称性:$M((F,E),(G,E)) = M((G,E),(F,E))$;

准则 3　归一性:$M((F,E),(G,E)) = 1 \Leftrightarrow (F,E) = (G,E)$;

准则 4　单调性:$(F,E) \subseteq (G,E) \subseteq (H,E)$,则:

$M((F,E), (H,E)) \leqslant \min(M((F,E), (H,E)), M((G,E),(H,E)))$。

通过 Vague 软集间相似度量的理化定义,可知 2 个 Vague 软集间的相似度量越大,则这 2 个 Vague 软集越相似。

2　新的 Vague 软集的相似度量

针对已有文献提出的 Vague 软集间相似度量的局限性,下面提出一种新的 Vague 软集间相似度量公式,充分考虑了 Vague 集的真隶属度、假隶属度、犹豫度、核以及 Vague 值的区间中心等区间特征因素。

定理 1　设 $U = \{x_1, x_2, \cdots, x_n\}$ 是一个论域,$E = \{e_1, e_2, \cdots, e_m\}$ 是一个参数集,$VSS(U)$ 表示论域 U 上的 Vague 软集,已知 (F,E),$(G,E) \in VSS(U)$,则称下式为 Vague 软集的相似度量:

$$M((F,E),(G,E)) = \sum_{i=1}^{m} \lambda_i \left\{ 1 - \frac{1}{7n} \sum_{j=1}^{n} \left[\begin{array}{l} | t_{F(e_i)}(x_j) - t_{G(e_i)}(x_j) | + \\ | f_{F(e_i)}(x_j) - f_{G(e_i)}(x_j) | + \\ | \pi_{F(e_i)}(x_j) - \pi_{G(e_i)}(x_j) | + \\ | S_{F(e_i)}(x_j) - S_{G(e_i)}(x_j) | + \\ | \phi_{F(e_i)}(x_j) - \phi_{G(e_i)}(x_j) | \end{array} \right] \right\}$$

式中,$\pi_{F(e_i)}(x_j) = 1 - t_{F(e_i)}(x_j) - f_{F(e_i)}(x_j)$ 和 $\pi_{G(e_i)}(x_j) = 1 - t_{G(e_i)}(x_j) - f_{G(e_i)}(x_j)$ 分别为 2 个 Vague 软集 $F(e_i)$ 和 $G(e_i)$ 中元素 x_j 的犹豫度,它表征对于参数 e_i 来说,现有证据对元素 x_j 的弃权信息。$S_{F(e_i)}(x_j) = t_{F(e_i)}(x_j) - f_{F(e_i)}(x_j)$ 和 $S_{G(e_i)}(x_j) = t_{G(e_i)}(x_j) - f_{G(e_i)}(x_j)$ 分别为 2 个 Vague 软集 $F(e_i)$ 和

$G(e_i)$ 中元素 x_j 的核,它表征对于参数 e_i 来说,现有证据对元素 x_j 支持和反对 2 种力量的对比。$\phi_{F(e_i)}(x_j)=\dfrac{1-t_{F(e_i)}(x_j)+f_{F(e_i)}(x_j)}{2}$ 和 $\phi_{G(e_i)}(x_j)=\dfrac{1-t_{G(e_i)}(x_j)+f_{G(e_i)}(x_j)}{2}$ 分别为 2 个 Vague 软集 $F(e_i)$ 和 $G(e_i)$ 中元素 x_j 的区间中心。λ_i 为参数 e_i 的权重。

下面证明新的 Vague 软集间相似度量是否满足公理化定义。

证明:

(1) 易知:$\pi_{F(e_i)}(x_j)\in[-1,1]$,$\pi_{G(e_i)}(x_j)\in[-1,1]$,$S_{F(e_i)}(x_j)\in[-1,1]$,$S_{G(e_i)}(x_j)\in[-1,1]$,$\phi_{F(e_i)}(x_j)\in[0,1]$,$\phi_{G(e_i)}(x_j)\in[0,1]$,$|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|\leqslant 1$,$|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|\leqslant 1$,又:

$|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|\leqslant 2$,$|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|\leqslant 2$,$|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|\leqslant 1$。因此,

$0\leqslant[|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|+|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|+|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|+|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|+|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|]\leqslant 7$;所以,

$$0\leqslant 1-\frac{1}{7n}\sum_{j=1}^{n}\begin{bmatrix}|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|+\\|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|+\\|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|+\\|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|+\\|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|\end{bmatrix}\leqslant 1,$$

则

$$0\leqslant\sum_{i=1}^{m}\lambda_i\left\{1-\frac{1}{7n}\sum_{j=1}^{n}\begin{bmatrix}|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|+\\|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|+\\|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|+\\|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|+\\|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|\end{bmatrix}\right\}=$$

$\sum_{i=1}^{m}\lambda_i\cdot 1=1$,$0\leqslant M((F,E),(G,E))\leqslant 1$,有界性成立,即能满足准则(1)。

(2) 由于

$|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|+|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|+|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|+|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|+|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|=|t_{G(e_i)}(x_j)-t_{F(e_i)}(x_j)|+|f_{G(e_i)}(x_j)-f_{F(e_i)}(x_j)|+|\pi_{G(e_i)}(x_j)-\pi_{F(e_i)}(x_j)|+|S_{G(e_i)}(x_j)-S_{F(e_i)}(x_j)|+|\phi_{G(e_i)}(x_j)-\phi_{F(e_i)}(x_j)|$,故 $M((F,E),(G,E))=M((G,E),(F,E))$,对称性成立,即能满足准则(2)。

(3) 由于 $M((F,E),(G,E))=1$,故

$$\begin{bmatrix}|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|+\\|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|+\\|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|+\\|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|+\\|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|\end{bmatrix}=0$$

,所以,

$|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|=|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|=|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|=|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|=|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|=0$

故 $t_{F(e_i)}(x_j)=t_{G(e_i)}(x_j)$,$f_{F(e_i)}(x_j)=f_{G(e_i)}(x_j)$,$\pi_{F(e_i)}(x_j)=\pi_{G(e_i)}(x_j)$,即,归一性成立,即能满足准则(3)。

(4) 因为 $(F,E)\subseteq(G,E)\subseteq(H,E)$,所以,

$t_{F(e_i)}(x_j)\leqslant t_{G(e_i)}(x_j)\leqslant t_{H(e_i)}(x_j)$,$f_{F(e_i)}(x_j)\geqslant f_{G(e_i)}(x_j)\geqslant f_{H(e_i)}(x_j)$,则:

$|t_{F(e_i)}(x_j)-t_{H(e_i)}(x_j)|\geqslant|t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)|$,$|f_{F(e_i)}(x_j)-f_{H(e_i)}(x_j)|\geqslant|f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)|$

又:$S_{F(e_i)}(x_j)-S_{H(e_i)}(x_j)=t_{F(e_i)}(x_j)-t_{H(e_i)}(x_j)+f_{H(e_i)}(x_j)-f_{F(e_i)}(x_j)$,

$S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)=t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)+f_{G(e_i)}(x_j)-f_{F(e_i)}(x_j)$,于是,

$|S_{F(e_i)}(x_j)-S_{H(e_i)}(x_j)|\geqslant|S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)|$;

又:$\pi_{F(e_i)}(x_j)-\pi_{H(e_i)}(x_j)=t_{H(e_i)}(x_j)-t_{F(e_i)}(x_j)+f_{H(e_i)}(x_j)-f_{F(e_i)}(x_j)$,

$\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)=t_{G(e_i)}(x_j)-t_{F(e_i)}(x_j)+f_{G(e_i)}(x_j)-f_{F(e_i)}(x_j)$,于是,

$|\pi_{F(e_i)}(x_j)-\pi_{H(e_i)}(x_j)|\geqslant|\pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)|$;

又,$\phi_{F(e_i)}(x_j)-\phi_{H(e_i)}(x_j)=\dfrac{1}{2}[t_{H(e_i)}(x_j)-t_{F(e_i)}(x_j)+f_{F(e_i)}(x_j)-f_{H(e_i)}(x_j)]$,$\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)=\dfrac{1}{2}[t_{G(e_i)}(x_j)-t_{F(e_i)}(x_j)+f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)]$,于是,

$|\phi_{F(e_i)}(x_j)-\phi_{H(e_i)}(x_j)|\geqslant|\phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)|$。

综上:

$$1-\frac{1}{7n}\sum_{j=1}^{n}\begin{bmatrix}|\ t_{F(e_i)}(x_j)-t_{G(e_i)}(x_j)\ |+\\ |\ f_{F(e_i)}(x_j)-f_{G(e_i)}(x_j)\ |+\\ |\ \pi_{F(e_i)}(x_j)-\pi_{G(e_i)}(x_j)\ |+\\ |\ S_{F(e_i)}(x_j)-S_{G(e_i)}(x_j)\ |+\\ |\ \phi_{F(e_i)}(x_j)-\phi_{G(e_i)}(x_j)\ |\end{bmatrix}\geqslant$$

$$1-\frac{1}{7n}\sum_{j=1}^{n}\begin{bmatrix}|\ t_{F(e_i)}(x_j)-t_{H(e_i)}(x_j)\ |+\\ |\ f_{F(e_i)}(x_j)-f_{H(e_i)}(x_j)\ |+\\ |\ \pi_{F(e_i)}(x_j)-\pi_{H(e_i)}(x_j)\ |+\\ |\ S_{F(e_i)}(x_j)-S_{H(e_i)}(x_j)\ |+\\ |\ \phi_{F(e_i)}(x_j)-\phi_{H(e_i)}(x_j)\ |\end{bmatrix}$$

即 $M((F,E),(G,E))\geqslant M((F,E),(H,E))$。同理可得,$M((H,E),(G,E))\geqslant M((F,E),(H,E))$,所以,

$M((F,E),(H,E))\leqslant \min(M((F,E),(H,E)),M((G,E),(H,E)))$。单调性成立,即能满足准则(4)。

证毕。

3　基于 Vague 软集相似度量的网络舆情综合评判方法

如何基于数据挖掘关键技术,实现高效畅通网上舆情的发现、分析、评估、预警、处置和反馈机制,是众多舆情监管部门亟待解决的重要问题。对如何在传播扩散、民众关注、内容敏感性、信息透明度、响应速度等多维度下,对网络舆情事件的安全性进行评估,从而甄别出苗头性、敏感性、危害性较大的网络舆情信息,是其中的关键环节。

设某舆情监管部门拟对一组网络舆情事件的安全性进行评估,从中筛选出最安全的舆情事件,有 5 个舆情事件可供研判,分别记为 X_1,X_2,X_3,X_4,X_5,这 5 个事件所具有的特征均以参数集表示:

$E=\{e_1,e_2,e_3,e_4,e_5\}=\{$传播扩散快,政府响应快,民众关注度高,信息透明度高,内容敏感度高$\}$。各参数的权重分别为$\{0.21,0.32,0.15,0.13,0.19,\}$。设定论域 U 仅包含支持和反对 2 个元素,记为 $U=\{$支持,反对$\}$。依据实际情况,舆情专家对最安全的舆情事件给出 Vague 软集 (X,E) 的评价值如表 1 所示,专家给出 5 个舆情事件的 Vague 软集评价值如表 2~6 所示。

表 1　最安全事件 X 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0,0]	[1,1]	[0,0]	[1,1]	[0,0]
反对	[1,1]	[0,0]	[1,1]	[0,0]	[1,1]

表 2　事件 X_1 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0.2,0.8]	[0.5,0.6]	[0.7,0.7]	[0.5,0.9]	[0.3,0.7]
反对	[0.4,0.9]	[0.5,0.7]	[0.7,0.9]	[0.8,0.8]	[0.2,0.8]

表 3　事件 X_2 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0.4,0.6]	[0.5,0.8]	[0.7,0.9]	[0.8,0.9]	[0.1,0.5]
反对	[0.7,0.7]	[0.8,0.8]	[0.9,0.9]	[0.9,0.9]	[0.6,0.7]

表 4　事件 X_3 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0.6,1.0]	[0.7,1.0]	[0.3,0.9]	[0.1,0.9]	[0.1,1.0]
反对	[0.7,0.9]	[0.8,0.9]	[0.5,0.5]	[0.4,0.4]	[0.3,0.8]

表 5　事件 X_4 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0.2,0.9]	[0.7,0.7]	[0.8,0.8]	[0.9,0.9]	[0.9,0.9]
反对	[0.1,0.9]	[0.4,0.7]	[0.4,0.9]	[0.6,0.9]	[0.7,0.9]

表 6 事件 X_5 的 $VSS(U)$

(X,E)	e_1	e_2	e_3	e_4	e_5
支持	[0.5,0.6]	[0.6,0.9]	[0.7,0.9]	[0.7,0.9]	[0.4,0.5]
反对	[0.4,0.6]	[0.5,0.8]	[0.7,0.9]	[0.8,0.9]	[0.1,0.5]

依据新的 Vague 软集相似度量公式分别计算事件 X_1,X_2,X_3,X_4,X_5 与最安全事件 X 评价值的相似度,结果如表 7 所示:

表 7 相似度计算结果

$M((X_i,E),(X,E))$	$M((X_1,E),(X,E))$	$M((X_2,E),(X,E))$	$M((X_3,E),(X,E))$	$M((X_4,E),(X,E))$	$M((X_5,E),(X,E))$
计算结果	0.602	0.648	0.591	0.626	0.606

结果显示,事件 X_1,X_2,X_3,X_4,X_5 与最安全事件 X 评价值的相似度可按降序排列为:$M((X_2,E),(X,E)) > M((X_4,E),(X,E)) > M((X_5,E),(X,E)) > M((X_1,E),(X,E)) > M((X_3,E),(X,E))$,可以看出:

事件 X_2 与最安全事件 X 评价值的相似度最高,为0.648,说明事件 X_2 的评价值最接近最安全事件 X 的评价值,因此事件 X_2 可划分为安全事件范畴。实验表明,基于 Vague 软集相识度量的舆情综合评判分析方法在实际问题中是有效和实用的。

4 实验及结果分析

为验证新的 Vague 软集相似度量方法在大规模网络舆情数据集下的综合评判效果,本节基于 MapReduce 框架模型对基于 Vague 软集相似度量的聚类算法并行化以改进传统的 Vague 软集聚类算法,使其适应 MapReduce 并行编程模型,从而能够有效地解决海量数据下的 Vague 软集聚类问题,以达到综合评判的效果。对大规模网络舆情数据集的实验结果证明,基于改进 Vague 软集相似度量的聚类算法在正确率和加速比性能方面,均优于传统的 Vague 软集聚类算法。

4.1 实验环境和数据集

本实验在由 7 台计算机组成的集群上运行,实验采用了 Apache 基金会下的 Hadoop 分布式框架。将其中 1 台机器作为主节点即 NameNode (或 JobTracker)节点,其余 6 台机器作为从节点即 DataNode(或 TaskTracker) 节点。每台机器的硬件配置如下:CPU 型号为 Intel Xeon7420 四核 64 位处理器,支持虚拟化,频率为 2.13GHz,内存大小为 64G,硬盘大小为 6T,操作系统为 Ubuntu 13.10,锐捷 RG-S2928G-E 千兆交换机,开发工具和平台为 Eclipse 8.5、JDK 1.7、Hadoop 2.7.1。

实验数据采用某社情民意大数据平台采集的真实微博舆情数据。该平台通过约 200 台服务器群不间断对涉及 40 000 个全国、全球重点网站、论坛的 150 000 个站点,4 家国内外微博等数据实时采集。目前该数据集搜集了已覆盖超过 350 000 个采集点,超过 1 亿的微博博主信息,微博入库量 1 000 万条。实验拟对微博热点话题进行聚类研究以综合评判,分别从聚类的准确率 PRE 和查全率 REC 来分析聚类的质量和评判效果,从算法的加速比 Sp 来衡量基于 MapReduce 的分块模糊聚类并行化的性能和效果。

4.2 算法加速比分析

为了测试算法的性能,实验中分别随机选取 5 组数据集进行测试,分别包含 3 000、10 000、100 000 条、500 000 条、1 000 000 条微博数据,分别从规模性、多样性、高速性、价值性 4 个参数特征考虑微博的舆情特性,其权值为{0.29,0.31,0.18,0.22}。对每一组数据分别使用基于 MapReduce 的 Vague 软集相似度量的聚类算法运行 8 次,实验中算法的加速比分析如表 8 所示:

表 8 算法的加速比分析

数据集(条)	3 000	10 000	100 000	500 000	1 000 000
加速比	0.447	1.141	2.447	3.834	7.737

从实验结果可以看出,当数据集较小时,算法在 Hadoop 分布式框架下的运行时间比单机环境下长,

主要是因为 MapReduce 过程中数据集的划分和聚类结果合并花费了较多的时间;而随着数据量不断增大时,通过 MapReduce 并行化改造后的聚类算法在 Hadoop 分布式框架下的运行时间明显低于单机环境下的运行时间,数据量越大则并行计算的优势越明显,Hadoop 系统对大规模数据集的处理能力也越强。实验表明基于 MapReduce 的 Vague 软集聚类算法在对大规模数据处理时能够得到较好的加速比。

4.3　算法准确率和查全率分析

由于 Vague 软集聚类评判结果受 Vague 软集之间相似度阈值选取的影响,因此实验采用新的相似度量的多个不同阈值进行实验,对每个阈值分别求出聚类的平均准确率和平均查全率,结果表明基于 MapReduce 的 Vague 软集聚类算法在 5 组数据集上的平均准确率和查全率均高于传统 Vague 软集聚类算法。实验结果如表 9 所示。

表 9　算法的准确率及查全率比较

算法	评价指标	3 000	10 000	100 000	500 000	1 000 000
传统 Vague 软集聚类	准确率	0.88	0.75	0.65	0.59	0.53
	查全率	0.87	0.77	0.74	0.71	0.68
Vague 软集并行化聚类	准确率	0.91	0.83	0.79	0.71	0.66
	查全率	0.94	0.85	0.83	0.79	0.75

分析发现,当聚类数据集规模较小时,2 种算法的准确率和查全率基本都在 0.85 以上,但当数据样本逐渐增大时,传统 Vague 软集聚类算法所得到的准确率和查全率与基于 MapReduce 的并行化聚类算法有明显差异,这是由于当数据量增大时,数据集中会出现很多非球形的不规则的类簇,而传统 Vague 软集聚类算法对于非球形簇并没有很好的聚类效果。基于 MapReduce 的 Vague 软集并行化聚类算法所得到的准确率和查全率明显优于传统 Vague 软集聚类算法。

5　结　论

本文在研究已有 Vague 软集相似度量问题的基础上,分析了现有 Vague 软集相似度量方法的不足,将 Vague 集的区间中心这一 Vague 集的重要参数特征引入 Vague 软集相似度量方法中开展研究,从而提出了一种新的 Vague 软集相似度量算法,并给出了公理化证明。通过对大规模舆情数据集的综合评判实验结果表明,该方法是一种有效的基于 Vague 软集相似度量的网络舆情综合评判分析方法。Vague 软集数学模型为解决网络舆情分析等决策问题提供了良好的理论工具和数学模型,有较好的应用前景。

参考文献:

[1] Gau W L, Buehrer D J. Vague Sets[J]. IEEE Trans on Systems, Man, and Cybmetics, 1993, 23(2): 610-614

[2] Molodtsov D. Soft Set Theory-First Results[J]. Computers & Mathematics with Applications, 1999, 37: 19-31

[3] Wei X, Jian M, Shou W, et al. Vague Soft Sets and Their Properties[J]. Computers & Mathematics with Applications, 2010, 59(2): 787-794

[4] Ganeshsree S. Vague Soft Rings and Vague Soft Ideals[J]. International Journal of Pure and Applied Mathematics, 2012, 6(12): 557-572

[5] Yun Y, Young J, Jianming Z. Vague Soft Hemirings[J]. International Journal of Pure and Applied Mathematics, 2011, 62(1): 199-213

[6] Nasruddin H, Khaleed A. Vague Soft Expert Set Theory[J]. AIP Advances, 2013(1522): 953-958

[7] Alhazaymeh K. Generalized Vague Soft Set and Its Applications[J]. International Journal of Pure and Applied Mathematics, 2012, 77(3): 391-401

[8] Alhazaymeh K, Nasruddin H. Interval-Valued Vague Soft Sets and Its Application[J]. Advances in Fuzzy Systems, 2012, 2012 (15): 1077-1083

[9] Teng Y, Wang C. Multicriteria Fuzzy Decision-Making Method Based on Vague Soft Sets[J]. Computer Engineering and Applications, 2012, 48(10): 6-8

[10] 王昌. Vague 软集的相似度量及其应用[J]. 统计与决策, 2012, 350(2):115-117
Wang Chang. Similarity Measurement and Application of Vague Soft Sets[J]. Statistics and Decision Making, 2012,350(2): 115-117 (in Chinese)

[11] Chang W, An Q. Entropy, Similarity Measure and Distance Measure of Vague Soft Sets and Their Relations[J]. Information Sciences, 2013, 244(20):92-106

[12] 陈文, 余本功. 基于 Vague 软集的模糊群决策方法研究[J]. 计算机工程与应用, 2014, 50(7):104-107
Chen Wen, YU Bengong. Research on Fuzzy Group Decision Making Method Based on Vague Soft Set[J]. Computer Engineering and Applications, 2014, 50(7):104-107 (in Chinese)

[13] 刘庆,王昌. 基于 Vague 软集的投资决策方案优选方法研究[J]. 科技通报, 2015, 31(1):4-8
Liu Qing, Wang Chang. Research on Optimized Method of Investment Decision Program Based on Vague Soft Sets[J]. Bulletin of Science and Technology, 2015, 31(1):4-8 (in Chinese)

[14] 刘庆,王昌. 基于 Vague 软集相似度量的快速估算模型[J]. 河北大学学报:自然科学版, 2014, 34(5):460-474
Liu Qing, Wang Chang. Fast Estimation Model Based on Similarity Measures Between Vague Soft Sets[J]. Journal of Hebei University: Natural Science Edition, 2014, 34(5):460-474 (in Chinese)

[15] 彭新东,杨勇. 区间值模糊软集的信息测度及其聚类算法[J]. 计算机应用, 2015,35(8):2350-2354
Peng Xindong, Yang Yong. Information Measures for Interval-Valued Fuzzy Soft Sets and Their Clustering Algorithm[J]. Journal of Computer Applications, 2015, 35(8):2350-2354 (in Chinese)

A Comprehensive Evaluation Method of Network Public Opinion Based On Similarity Measure Between Vague Soft Sets

Wang Wei[1], Wu Junsheng[1], Zhu Zhixiang[2], Yang Wenchao[3]

(1.School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 710072, China;
2.Institute of Internet of Things & Integration of Information and Industrialization,
University of Posts and Telecommunication, Xi'an 710061, China;
3.School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Aiming at the limitation of existing similarity measure method based on Vague soft sets, the similarity measure formula between Vague soft sets is modified and a novel similarity measure between Vague soft sets in consideration of the difference of interval center of Vague values is introduced, the axiomatic proof is given too. The experimental results of comprehensive evaluation of the network public opinion show that this method is reasonable, effective and practical, which has a good application prospect and effect in the study of internet public opinion.

Keywords: Vague soft sets; similarity measure; internet public opinion