基于直觉模糊推理的网络舆情预警方法 *

李弼程, 王 瑾, 林 琛
(解放军信息工程大学 信息工程学院, 郑州 450002)

摘 要: 为解决网络舆情预警等级问题,提出了一种基于直觉模糊推理的网络舆情预警方法。借鉴战场态势分析思想,对网络舆情态势分析的原理进行了阐述,选取了适合计算机实现的七个网络舆情态势分析模式对预警等级进行判断。选择七个舆情话题进行实验,实验结果表明,该方法能够准确地估计出威胁等级,符合专家经验判断,说明该方法是可行的。

关键词: 网络舆情; 态势分析; 威胁估计; 预警; 直觉模糊推理
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)09-3312-04
doi:10.3969/j.issn.1001-3695.2010.09.029

Method of online public opinions pre-warning based on intuitionistic fuzzy reasoning

LI Bi-cheng, WANG Jin, LIN Chen
(*Information Engineering Institute, Information Engineering University of PLA, Zhengzhou* 450002, *China*)

Abstract: To the issues of online public opinions pre-warning, this paper proposed a pre-warning method based on intuitionistic fuzzy reasoning. Described the theory of the state and trend analysis about online public opinions, which came from the battlefield state and trend analysis in the field of military, and it selected seven modes of the state and trend analysis about online public opinions. The experimental result using 7 topics of online public opinions shows that the method can estimate the threat's level much precisely, and to a degree, the estimate is fit for the opinion of experts' experience. So, the method is availiable.

Key words: online public opinions; state and trend analysis; threat estimating; pre-warning; intuitionistic fuzzy reasoning

网络舆情是指在互联网上传播的公众对某一关注事件或话题所表现的有一定影响力、具有倾向性的意见或言论的情况,是社会舆情的一种重要表现形式。其表达快捷、信息多元、方式互动等特点使得其形成速度快,对社会影响越来越巨大。因此,党和政府需要对网络舆情进行及时的掌控。

近年来,网络舆情分析与预警已经成为研究的热点,其中网络舆情态势分析是预警的前提,否则无法进行有效的预警。谢海光等人[2]构建了互联网内容与舆情的热点(热度)、重点(重度)、焦点(焦度)、敏点(敏度)等十个分析模式和判据。钱爱兵[2]设计出主题关注度分析、热点分析、焦点分析、拐点分析、重点分析。吴绍忠等人[3]研究了网络舆情预警机制,设立网络舆情预警等级,从定性方面设计网络舆情预警指标体系,运用 Delphi 法确定指标体系权重。目前,对舆情的预警等级评估的研究非常少,没有相关的评估模型和计算方法,大都靠人工进行判断,不能全面、及时、准确地判断舆情威胁。

在网络舆情预警分析系统开发方面,比较有代表性是北大方正技术研究院的方正智思舆情预警辅助决策支持系统[4]、Autonomy[5]的互联网舆情监控分析系统、Goonie 网络舆情监控分析系统[6]和 TRS 网络舆情监控系统[7]等。这些系统大多利用话题检测与追踪、文本检索等领域的工具对网络舆情进行分析,主要注重获取舆情话题的主题内容,忽视了公众与话题之间的关系、同一话题中不同事件之间的相互关系以及这些关系的变化趋势。预警信息也只是利用统计方法通过简单的报告、图表的形式给出话题变化趋势,缺乏舆情预警等级评估这一环节。

1 网络舆情态势分析与预警的基本原理

通常,态势分析和威胁估计是两个军事术语,即战场的态势分析和威胁估计[8]。本文把军事领域中的战场态势分析和威胁估计的思想引入到网络舆情分析和预警中。经过类比分析,得到网络舆情和战场的对应关系为:舆情中的话题(事件)对应于战场的目标;舆情中的公众对应于战场的环境。网络舆情预警就是根据网络舆情态势信息,利用威胁估计技术对网络舆情的威胁程度进行定量估计,作出网络舆情的预警等级预报。其中,网络舆情的预警等级[3]可以划分为轻警情(Ⅳ级,非常态)、中度警情(Ⅲ级,警示级)、重警情(Ⅱ级,危险级)和特重警情(Ⅰ级,极度危险级)四个等级,并依次采用蓝色、黄色、橙色和红色来加以表示。

a)蓝色级(Ⅳ级)。网民对该舆情关注度低,传播速度慢,没有转换为行为舆论的可能。

b)黄色级(Ⅲ级)。网民对该舆情关注度较高,传播速度中等,没有转换为行为舆论的可能。

收稿日期: 2010-03-22; **修回日期**: 2010-04-12 **基金项目**: 国家"863"计划资助项目(2007AA01Z439)

作者简介:李弼程(1970-),男,湖南衡南人,教授,博导,博士,主要研究方向为智能信息处理;王瑾(1981-),男,湖南新化人,硕士研究生,主要研究方向为信息融合(ljjg_1005@163.com);林琛(1981-),女,山东文登人,助理工程师,博士研究生,主要研究方向为网络舆情分析.

c)橙色级(Ⅱ级)。网民对该舆情关注度高,传播速度快,舆情有转换为行为舆论的可能。

d)红色级(Ⅰ级)。网民对该舆情关注度极高,传播速度非常快,舆情即将转换为行为舆论。

目前战场威胁估计的主要方法有多属性决策[9]、神经网络方法[10]、直觉模糊推理[11,12]、支持向量机[13]、模糊集合论[14,15]等,主要集中在战场态势和目标的威胁程度判断方面。由于态势分析和威胁估计涉及相关领域的诸多背景知识,具有多方面的不确定性。总体而言,研究难度较大,仍是当前信息融合的薄弱环节。

2　网络舆情态势分析模式

根据上述网络舆情态势分析与预警的基本原理,本文选择了话题、公众、话题与公众之间的关系三大类,共七个相对独立的、适合计算机实现的网络舆情预警等级分析模式,如图 1 所示。

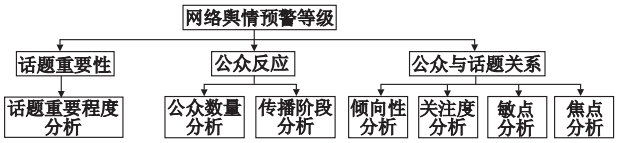

图1　网络舆情预警等级分析模式

2.1　话题分析模式

话题分析模式主要是话题重要性分析,分析舆情话题内容的重要程度,对各个舆情话题内容的重要程度进行归一化,得到归一化重要程度 $x_1 \in [0,1]$。一般地,x_1 越高,危险级别越高。

2.2　公众分析模式

a)公众数量分析,就是估计舆情受众的数量,其实现技术是对点击率进行统计。受众的数量越多,舆情的影响力就越大。对舆情受众的数量进行归一化,得到归一化公众数量 $x_2 \in [0,1]$。x_2 越高,危险级别越高。

b)传播阶段分析。传播速度分析是分析舆情受众的数量的变化速度,其实现技术是统计点击率的变化。对传播加速度进行分段线性映射,得到归一化传播加速度 $x_3 \in [0,1]$。其中:$x_3 \in [0,0.5)$ 表示消退阶段,$x_3 = 0.5$ 表示稳定阶段;$x_3 \in (0.5,1]$ 表示扩散阶段。x_3 越高,危险级别越高。

2.3　话题与公众之间的关系分析模式

1)倾向性分析　就是提取公众对舆情话题所表现的情感信息,评估民众对舆情话题的态度倾向性,包括强烈支持、支持、中立、反对、强烈反对。针对对立面的态度倾向性越强烈,危险级别较高。对舆情话题的态度倾向性程度进行归一化,得到归一化倾向性程度 $x_4 \in [0,1]$。其中,$x_4 = 0.5$ 表示强烈反对所处的对立面;$x_4 = 0$ 表示中立;$x_4 = 1$ 表示强烈支持对立面。x_4 越高,危险级别越高。例如祖国 60 周年庆典相关报道中,设定对立面为负面情绪,则当倾向性为强烈支持时,危险级别较低。

2)关注度分析　就是估计在过去某一时间段内舆情话题被关注的程度,一般用该舆情话题的相关网页数进行衡量[2]。对舆情关注度进行归一化,得到归一化关注度 $x_5 \in [0,1]$。x_5 越高,危险级别越高。

3)敏点分析　识别某一时间段内在热点排行榜上位次上升较多的舆情话题。其实现技术是统计各舆情话题在热点排行表的位次上升情况[1],用上升幅度来表示,上升幅度大于给定阈值的舆情话题为敏点。一般地,上升幅度越大,危险级别较高。对上升幅度进行归一化,得到归一化上升幅度 $x_6 \in [0,1]$。x_6 越高,危险级别越高。

4)焦点分析　识别持续的热点舆情话题。其实现技术是统计在持续数个数据统计期内在热点排行榜上保持较高位次的舆情话题[1]。一般地,持续时间越长,危险级别较高。对持续时间进行归一化,得到归一化持续时间 $x_7 \in [0,1]$。x_7 越高,危险级别越高。

3　基于直觉模糊推理的网络舆情预警等级评估

Atanassov[17,18]的直觉模糊集理论是对 Zadeh 模糊集理论最具影响力的一种推广,且其数学描述较之 Zadeh 模糊集理论更加符合客观世界模糊对象的本质。因此,本文采用直觉模糊推理进行舆情的威胁等级评估。基于直觉模糊推理的网络舆情预警等级评估基本思想如下:

a)将上述七个网络舆情态势分析模式转换为区间 $[0,1]$ 的度量 $x_1, x_2, \cdots, x_7$,基本方法是线性映射或分段线性映射。

b)通过直觉模糊综合评判的方法得出每个因素的隶属度。参与推理的要素因素包括话题重要性、公众反应、公众与话题之间联系,其论域分别为话题重要性论域、公众反应论域、公众与话题之间联系论域。

c)利用直觉模糊推理判断网络舆情预警等级,包括轻警情(Ⅳ级,非常态)、中度警情(Ⅲ级,警示级)、重警情(Ⅱ级,危险级)和特重警情(Ⅰ级,极度危险级)四个等级。

3.1　舆情要素及舆情威胁等级直觉模糊集

论域 X 上的直觉模糊集是下列形式的一个对象:

$$A = \{(x, \mu_A(x), \gamma_A(x)) \mid x \in X\}$$

其中:$\mu_A(x): X \to [0,1]$ 为 x 属于 A 的隶属度,$\gamma_A(x): X \to [0,1]$ 为 x 不属于 A 的隶属度,其和满足:

$$0 \leqslant \mu_A(x) + \gamma_A(x) \leqslant 1, \forall x \in A。$$

对于 X 上的直觉模糊集,定义 x 的直觉指数为

$$\pi_A(x) = 1 - \mu_A(x) - \gamma_A(x)$$

它反映了 x 对 A 犹豫程度的一种测度。

直觉模糊集 A 可以简记为 $A = \langle x, \mu_A, \gamma_A \rangle$,或 $A = \langle \mu_A, \gamma_A \rangle / x$。直觉模糊集 A 的补集 $\overline{A}$ 定义为

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x), \gamma_{\overline{A}}(x) = 1 - \gamma_A(x)$$

直觉模糊推理的前提是根据观测数据的物理意义,建立相应的隶属度函数和非隶属度函数,将观测数据进行模糊化。各直觉模糊子集的直觉模糊隶属度可以根据模式分析的结果进行赋值。

3.1.1　话题重要性直觉模糊集

话题重要性可以通过话题重要程度分析获得。取话题重要性论域 $A = [0,1]$,$x_1 \in [0,1]$ 为归一化重要程度,直觉模糊子集 $A_1 = \{\text{一般}\}$ 与 $A_2 = \{\text{重要}\}$ 的直觉模糊隶属度函数及直觉指数为

$$\mu_{A_1}(x_1) = 0.9 \exp\left(-\frac{x_1^2}{2\sigma_\mu^2}\right),\ \mu_{A_2}(x_1) = 0.9 \exp\left(-\frac{(x_1-1)^2}{2\sigma_\mu^2}\right)$$

$$\pi_{A_1}(x_1)=0.1\exp\left(-\frac{(1/2-x_1)^2}{2\sigma_\pi^2}\right),\ \pi_{A_2}(x_1)=0$$

其中:σ_μ 和 σ_π 为宽度。本文取 $\sigma_\mu=0.1416,\sigma_\pi=0.035$。话题重要性的直觉模糊集子集表示为$\langle x_1,\mu_{A_1},\gamma_{A_1}\rangle,\langle x_1,\mu_{A_2},\gamma_{A_2}\rangle$。

3.1.2　公众反应直觉模糊集赋值

公众反应可以通过公众数量分析、传播速度分析的结果融合获得,利用直觉模糊综合评判加权平均作为公众反应的直觉模糊集的隶属度。

设 $x_2\in[0,1]$ 为归一化公众数量,$x_3\in[0,1]$ 为归一化传播速度,公众反应论域 $B=[0,1]\times[0,1]$,取其直觉模糊子集为 $B_1=\{$慢$\}$,$B_2=\{$快$\}$,可得到公众反应的直觉模糊集子集表示为$\langle(x_2,x_3),\mu_{B_1},\gamma_{B_1}\rangle,\langle(x_2,x_3),\mu_{B_2},\gamma_{B_2}\rangle$。直觉模糊隶属度函数与直觉指数定义为

$$\mu_{B_1}(x_2,x_3)=0.45\left[\exp\left(-\frac{x_2^2}{2\sigma_\mu^2}\right)+\exp\left(-\frac{x_3^2}{2\sigma_\mu^2}\right)\right]$$

$$\mu_{B_2}(x_2,x_3)=0.45\left[\exp\left(-\frac{(1-x_2)^2}{2\sigma_\mu^2}\right)+\exp\left(-\frac{(1-x_3)^2}{2\sigma_\mu^2}\right)\right]$$

$$\pi_{B_1}(x_2,x_3)=0.05\left[\exp\left(\frac{(1/2-x_2)^2}{2\sigma_\pi^2}\right)+\exp\left(\frac{(1/2-x_3)^2}{2\sigma_\pi^2}\right)\right]$$

$$\pi_{B_2}(x_2,x_3)=0$$

其中:σ_μ 和 σ_π 为宽度。本文取 $\sigma_\pi=0.1416,\sigma_\pi=0.035$。

3.1.3　话题与公众联系直觉模糊集赋值

话题与公众之间的联系则可以通过倾向性分析、关注度分析、敏点分析和焦点分析的结果融合获得,同样利用直觉模糊综合评判加权平均作为话题与公众联系的直觉模糊集的隶属度。设 $x_4\in[0,1]$ 为归一化倾向性程度,$x_5\in[0,1]$ 为归一化关注度,$x_6\in[0,1]$ 为归一化上升幅度,$x_7\in[0,1]$ 为归一化持续时间;公众与话题联系论域 $C=[0,1]^4$,取其直觉模糊子集为 $C_1=\{$稀疏$\}$,$C_2=\{$紧密$\}$,可得到公众反应的直觉模糊集子集表示为$\langle(x_4,x_5,x_6,x_7),\mu_{C_1},\gamma_{C_1}\rangle,\langle(x_4,x_5,x_6,x_7),\mu_{C_2},\gamma_{C_2}\rangle$。直觉模糊隶属度函数与直觉指数定义为

$$\mu_{C_1}(x_4,x_5,x_6,x_7)=0.225[\exp\left(-\frac{x_4^2}{2\sigma^2}\right)+\exp\left(-\frac{x_5^2}{2\sigma^2}\right)+\exp\left(-\frac{x_6^2}{2\sigma^2}\right)+\exp\left(-\frac{x_7^2}{2\sigma^2}\right)]$$

$$\mu_{C_2}(x_4,x_5,x_6,x_7)=0.225[\exp\left(-\frac{(1-x_4)^2}{2\sigma^2}\right)+\exp\left(-\frac{1-x_5)^2}{2\sigma^2}\right)+\exp\left(-\frac{(1-x_6)^2}{2\sigma^2}\right)+\exp\left(-\frac{(1-x_7)^2}{2\sigma^2}\right)]$$

$$\pi_{C_1}(x_4,x_5,x_6,x_7)=0.025[\exp\left(-\frac{(0.5-x_4)^2}{2\sigma^2}\right)+\exp\left(-\frac{(0.5-x_5)^2}{2\sigma^2}\right)+\exp\left(-\frac{(0.5-x_6)^2}{2\sigma^2}\right)+\exp\left(-\frac{(0.5-x_7)^2}{2\sigma^2}\right)]$$

$$\pi_{C_2}(x_4,x_5,x_6,x_7)=0$$

其中:σ_μ 和 σ_π 为宽度。本文取 $\sigma_\mu=0.1416,\sigma_\pi=0.035$。

3.1.4　网络舆情预警等级直觉模糊隶属度

网络舆情预警等级论域 $Z=[0,1]$,取其直觉模糊子集为 $Z_1=\{$蓝色级$\}$,$Z_2=\{$黄色级$\}$,$Z_3=\{$橙色级$\}$,$Z_4=\{$红色级$\}$。直觉模糊隶属度函数与直觉指数为

$$\mu_{Z_1}(z)=0.9\exp\left(-\frac{z^2}{2\sigma^2}\right),\ \mu_{Z_2}(z)=0.9\exp\left(-\frac{(z-1/3)^2}{2\sigma^2}\right)$$

$$\mu_{Z_3}(z)=0.9\exp\left(-\frac{(z-2/3)^2}{2\sigma^2}\right),\ \mu_{Z_4}(z)=0.9\exp\left(-\frac{(z-1)^2}{2\sigma^2}\right)$$

$$\pi_{Z_1}(z)=0.1\exp\left(-\frac{(1/6-z)^2}{2\sigma^2}\right),\pi_{Z_2}(z)=0.1\exp\left(-\frac{(1/2-z)^2}{2\sigma^2}\right)$$

$$\pi_{Z_3}(z)=0.1\exp\left(-\frac{(5/6-z)^2}{2\sigma^2}\right),\ \pi_{Z_4}(z)=0$$

其中:$z\in[0,1]$,σ_μ 和 σ_π 为宽度。本文取 $\sigma_\mu=0.07,\sigma_\pi=0.035$。

以上各分析模式以及舆情等级选取的直觉模糊隶属度函数参数,对能够提取出上述七种模式的新闻和论坛话题均有很好的适用性。

3.2　直觉模糊推理规则

网络舆情预警等级的输入参数等变量的子集个数分别为 $N_a=2,N_b=2,N_c=2$;输出个数 $N_z=4$。则由以上可计算出系统中推理规则的个数为 $N=N_a\times N_b\times N_c=8$。

简单的规则产生式形如:if E then H;复杂条件可以在规则中使用 and、or 等,即使用组合规则。本文中采用的条件有三个,因而用“and”操作连接模糊推理形式为

$R^{(k)}$:if a is A_{ia} and b is B_{ib} and c is C_{ic} then z is Z_{jz}

其中:$k=1,2,\cdots,8$;$ia=1,2$;$ib=1,2$;$ic=1,2$,$jz=1,2,3,4$;a、b、c 为输入变量,而 z 为输出变量;A_{ia},B_{ib},C_{ic} 为前提部分语言项,分别为$\langle a,\mu_{ia},\gamma_{ia}\rangle$,$a=x_1\in A$;$\langle b,\mu_{ib},\gamma_{ib}\rangle$,$b=(x_2,x_3)\in B$;$\langle c,\mu_{ic},\gamma_{ic}\rangle$,$c=(x_4,x_5,x_6,x_7)\in C$,$Z_{jz}$ 为输出论域中的一个模糊子集 Z_m,即$\langle z,\mu_{jz},\gamma_{jz}\rangle$,$z\in Z$,$m=1,2,3,4$。

根据专家经验,建立规则如下:

$R^{(1)}$:if a is A_1 and b is B_1 and c is C_1 then z is Z_1

$R^{(2)}$:if a is A_2 and b is B_1 and c is C_1 then z is Z_1

$R^{(3)}$:if a is A_1 and b is B_2 and c is C_1 then z is Z_2

$R^{(4)}$:if a is A_1 and b is B_1 and c is C_2 then z is Z_2

$R^{(5)}$:if a is A_2 and b is B_2 and c is C_1 then z is Z_3

$R^{(6)}$:if a is A_2 and b is B_1 and c is C_2 then z is Z_3

$R^{(7)}$:if a is A_1 and b is B_2 and c is C_2 then z is Z_4

$R^{(8)}$:if a is A_2 and b is B_2 and c is C_2 then z is Z_4

每一条规则 $R^{(k)}$ 都是一个单值输出,本文采取“$\wedge$ – $\vee$”合成运算。令 $G=A_{ia}\cap B_{ib}\cap C_{ic}$,满足:

$$\mu_G(x_1,x_2,x_3,x_4,x_5,x_6,x_7)=\mu_{A_{ia}}(x_1)\wedge\mu_{B_{ib}}(x_2,x_3)\wedge\mu_{C_{ic}}(x_4,x_5,x_6,x_7)$$

$$\gamma_G(x_1,x_2,x_3,x_4,x_5,x_6,x_7)=\gamma_{A_{ia}}(x_1)\vee\gamma_{B_{id}}(x_2,x_3)\vee\gamma_{C_{ic}}(x_4,x_5,x_6,x_7)$$

直觉模糊关系

$$R^{(k)}(A_{ia}\cap B_{ib}\cap C_{ic}\rightarrow Z_{jz})=R^{(k)}(A_{ia},B_{ib},C_{ic};Z_{jz})$$

满足:

$$\mu_{R(k)}(x_1,x_2,x_3,x_4,x_5,x_6,x_7;z)=\mu_G(x_1,x_2,x_3,x_4,x_5,x_6,x_7)\wedge\mu_{Z_{jz}}(z)$$

$$\gamma_{R(k)}(x_1,x_2,x_3,x_4,x_5,x_6,x_7;z)=\gamma_G(x_1,x_2,x_3,x_4,x_5,x_6,x_7)\vee\gamma_{Z_{jz}}(z)$$

当某一时刻输入七个网络舆情态势分析模式的归一化度量 $x'_1,x'_2,\cdots,x'_7$,先进行模糊化,即定义直觉模糊集 A'、B'和 C':

$$\mu_{A'}(x_1)=\exp\left(-\frac{(x_1-x'_1)^2}{2\sigma^2}\right)$$

$$\mu_{B'}(x_2,x_3)=\frac{1}{2}\left[\exp\left(-\frac{(x_2-x'_2)^2}{2\sigma^2}\right)+\exp\left(-\frac{(x_3-x'_3)^2}{2\sigma^2}\right)\right]$$

$$\mu_C(x_4,x_5,x_6,x_7)=\frac{1}{4}\left[\exp\left(\frac{(x_4-x'_4)^2}{2\sigma^2}\right)+\exp\left(\frac{(x_5-x'_5)^2}{2\sigma^2}\right)+\exp\left(\frac{(x_6-x'_6)^2}{2\sigma^2}\right)+\exp\left(\frac{(x_7-x'_7)^2}{2\sigma^2}\right)\right]$$

直觉模糊集 A'、B'和 C'的直觉指数为 0,不影响方法的有效性。

令 $G'=A'\times B'\times C'$,由推理规则可以得出 $Z'_k=G'\circ R^{(k)}$,满足:

$$\mu_{Z'_k}(z)=\bigvee_{x_1,x_2,\cdots,x_7}(\mu_{G'}(x_1,x_2,\cdots,x_7)\wedge\mu_{R(k)}(x_1,x_2,\cdots,x_7;z))$$

$$\gamma_{Z'_k}(z)=\bigwedge_{x_1,x_2,\cdots,x_7}(\gamma_{G'}(x_1,x_2,\cdots,x_7)\vee\gamma_{R(k)}(x_1,x_2,\cdots,x_7;z))$$

其中:$x_1,x_2,\cdots,x_7,z\in[0,1]$。

合成直觉模糊集 $Z'_k,k=1,2,\cdots,8$,得到直觉模糊集 Z':

$$\mu_{Z'}(z)=\bigvee_k\mu_{Z'_k}(z),\gamma_{Z'}(z)=\bigwedge_k\gamma_{Z'_k}(z)$$

3.3　判决准则

利用直觉模糊集 Z'估计网络舆情预警等级,需要利用直觉模糊集之间接近程度(即贴近度)进行模式分类[16]。本文采用直觉模糊的真值来计算直觉模糊集之间的贴近度。

论域 Z 上的直觉模糊集 Q 的真值为

$$T_Q(z)=\alpha\cdot\mu_Q(z)+\beta\cdot\pi_Q(z),z\in Z$$

一般取 $\alpha=1,\beta=1/2$。论域 Z 上的直觉模糊集 P 与 Q 之间的贴近度 $\sigma(\cdot,\cdot)$定义为

$$\sigma(P,Q)=1-\left\{\int_0^1[T_P(u)-T_Q(u)]^2\mathrm{d}u\right\}^{1/2}$$

对于论域 $Z=[0,1]$的直觉模糊集 $Z_1=\{$蓝色级$\}$,$Z_2=\{$黄色级$\}$,$Z_3=\{$橙色级$\}$,$Z_4=\{$红色级$\}$,如果存在 $i\in\{1,2,3,4\}$,有 $\sigma(Z',Z_i)=\max\limits_{1\leqslant j\leqslant 4}\sigma(Z',Z_j)$。则称 Z'与 Z_i 最贴近,并把 Z_i 作为网络舆情预警等级。

4　实验结果分析

从 2009 年百度贴吧以及网易新闻的跟帖中采集到的舆情话题中选取七个舆情话题,分别是网易新闻跟帖:(1)国庆 60 周年报道专题“我爱我的中国”;(5)哥本哈根气候变化大会;(6)河南申论考 0 分进入公务员面试线;(7)南京安德门民工猝死;百度贴吧事件:(2)周久耕事件,(3)范美忠事件,(4)三鹿奶粉事件。对这些话题采集数据分别进行分析并归一化,如表 1 所示,通过推理得出各话题网络舆情预警等级如表 2 所示。

表 1　舆情数据检测系统检测数据归一化结果

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7
(1)	0.22	0.94	0.51	0.01	0.70	0.68	0.13
(2)	0.65	0.73	0.69	0.96	0.62	0.59	0.37
(3)	0.47	0.85	0.41	0.64	0.35	0.41	0.48
(4)	0.59	0.87	0.38	0.77	0.37	0.36	0.29
(5)	0.31	0.23	0.65	0.52	0.21	0.22	0.27
(6)	0.32	0.18	0.44	0.63	0.18	0.15	0.13
(7)	0.45	0.21	0.32	0.76	0.11	0.16	0.15

表 2　推理结果及网络舆情等级判断结果

序号	σ_1	σ_2	σ_3	σ_4	等级
(1)	0.6615	0.8155	0.6025	0.7356	Ⅲ级
(2)	0.7017	0.6642	0.7353	0.8134	Ⅰ级
(3)	0.7532	0.7347	0.7092	0.7516	Ⅳ级
(4)	0.7289	0.6575	0.7774	0.7376	Ⅱ级
(5)	0.8069	0.7498	0.6332	0.6852	Ⅳ级
(6)	0.8530	0.7036	0.6417	0.6946	Ⅳ级
(7)	0.7714	0.7557	0.6978	0.7057	Ⅳ级

例 1　对第二个话题的检测值,输入向量为(0.65,0.73,0.69,0.96,0.62,0.59,0.37),对向量模糊化,经过推理可得输出的贴近度向量为(0.7017,0.6642,0.7353,0.8134),判断网络舆情预警等级为Ⅰ级。

例 2　对第四个话题的检测值,输入向量为(0.59,0.87,0.38,0.77,0.37,0.36,0.29),对向量模糊化,经过推理可得输出贴近度向量为(0.7289,0.6575,0.7774,0.7376),判断网络舆情预警等级为Ⅱ级。

对照表 1 和 2 可以看出,网络舆情预警等级估计的结果是准确的,符合专家经验判断。

5　结束语

近年来,网络舆情分析与预警已经成为研究的热点,但是,对舆情的预警等级评估的研究非常少,没有相关的评估模型和计算方法,大都靠人工进行判断,不能全面、及时、准确地判断舆情威胁。本文利用军事领域中的战场态势分析思想阐述了网络舆情态势分析的原理,构建了适合计算机实现的网络舆情态势分析模式,在此基础上,采用直觉模糊推理技术,为利用计算机自动判断网络舆情预警等级提供一定思路,实验结果验证了本文方法的有效性。

参考文献:

[1] 谢海光, 陈中润. 互联网内容及舆情深度分析模式[J]. 中国青年政治学院学报, 2006, 3:95-100.

[2] 钱爱兵. 基于主题的网络舆情分析模型及其实现[J]. 现代图书情报技术, 2008, 4:49-55.

[3] 吴绍忠, 李淑华. 互联网络舆情预警机制研究[J]. 中国人民公安大学学报:自然科学版, 2008, 3(3):38-42.

[4] 方正智思舆情预警辅助决策支持系统[EB/OL]. [2006-03-16]. http://www.founderrd.com/2006-03/16.

[5] Autonomy 面向中国市场推出三款企业搜索产品[EB/OL]. [2006-10-10]. http://news.csdn.net/n/20061010/95941.html.

[6] Goonie 网络舆情监控分析系统[EB/OL]. [2008-01-11]. http://www.goonie.cn/products/2008/01/content3.html.

[7] 都云程,王海洋,王洪俊. TRS 网络舆情监控解决方案[J]. 信息网络安全, 2008(6):69-70.

[8] 刘同明,夏祖勋,解洪威. 数据融合技术及其应用[M]. 北京:清华大学出版社, 1998: 230-236.

[9] 柯宏发, 陈永光. 电子战干扰目标的多属性多层次威胁评估模型[J]. 系统工程与电子技术, 2006, 28(9):1370-1374.

[10] 王向华,覃征,刘宇,等. 径向基神经网络解决威胁排序问题[J]. 系统仿真学报, 2004, 16(7):1576-1579.

[11] 雷英杰,王宝树,王毅. 基于直觉模糊推理的威胁评估方法[J]. 电子与信息学报, 2007, 29(9):2077-2081.

[12] 雷英杰,王宝树,路艳丽. 基于自适应直觉模糊推理的威胁评估方法[J]. 电子与信息学报, 2007, 29(12):2805-2809.

[13] 袁斌,耿伯英,杨红梅. 基于支持向量机的海战场辐射源威胁评估[J]. 火力与指挥控制, 2008, 33(2):63-65.

[14] GONSALVES P, CUNNINGHAM R, TON N, *et al*. Intelligent threat assessment processor(ITAP) using genetic algorithms and fuzzy logic[C]//Proc of the 3rd International Conference on Information Fusion. 2000:18-24.

(下转第 3325 页)

实验结果证明了 2.3 节中的两个分析:a)随着计算块块数的增多证认计算量逐渐减少,由此带来了证认总耗时的减少;b)I/O 读取的耗时、随着副本数量增多带来的额外数据处理的耗时会平抑距离计算量减少对整体效率的影响,所以总耗时曲线的下降幅度明显小于图 3 中球面距离计算量理论上的下降幅度。在存储量方面,实验结果与 2.3 节中理论上分析的结果基本一致。由此可知,分块粒度取到 12×4^{10} 或 12×4^{11} 比较合理,在没有引入过多存储量的同时,证认效率基本达到最高。

实验 2　随节点个数的增加证认效率加速比的测试

对于分布式程序,加速比的高低决定了算法能否具有良好的规模扩展性。本文将整个交叉证认过程分成数据的分布式存放和证认计算两个层次,从而将影响加速比提升的数据通信部分尽可能地解决在数据的分布式存放阶段,以求获得最好的证认计算性能。在 4、8、16、32、64 节点集群上、分块粒度下,证认计算部分的性能结果如图 6 所示。

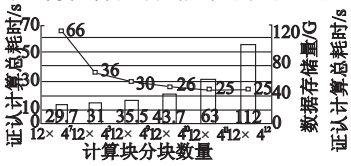

图5　证认计算耗时和数据存储量随分块数量的变化

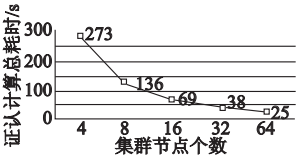

图6　不同节点个数下证认计算总耗时的变化情况

可以看出,从 4 ~ 32 节点,基本上达到了线性加速,即节点数每扩大一倍,时间约减少一半;而当节点数增加到 64 时,效率提高 30%。这与 Hadoop 计算过程有一定的启动时间有关,当证认时间缩减到一定程度时,启动时间占有的比例相应提高,致整体加速比减少。接近线性的加速比使本文方法继续推广到更大数据集、更大集群规模成为了可能。考虑到用户在提交交叉证认请求时多是针对某一天区的,而非当前的全天区证认,所以本文方法的效率已经基本可以满足实时交叉证认服务的需求。

实验 3　数据的分布式存放部分性能测试

在构建交叉证认服务或多源联合查询服务时,只有证认计算部分是用户提交请求后的操作,故只有这部分关系到服务实时性的实现。数据的分布式存放可看做预处理过程,只在一个数据集新加入时执行一次,因此这部分的性能只要满足基本要求即可。图 7 给出的是这部分在不同节点数下的执行时间,可以看出其耗时也是完全可以接受的。

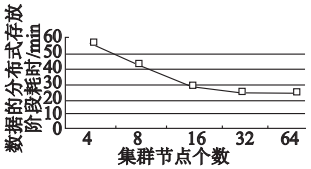

图7　证认数据的分布式存放部分耗时

实验 4　与其他方法性能对比

以上实验表明文中方法在上亿条的数据规模上确实表现出了很高的计算效率,25 s 的证认耗时相比先前多核环境下的并行方法[7]在完全相同数据集上的 32 min 的运行时间提升非常显著,而对比高丹等人[5]的方法的效率更是提高了万倍以上。

4　结束语

本文提出了一种基于 MapReduce 的天文交叉证认的分布式并行计算方法,将整个交叉证认过程分成了数据分发、证认计算两个独立的 MapReduce 过程。其中数据分发阶段是对数据的预处理,在交叉证认平台中只需对新加入的数据执行一次,所以设计上这部分尽量地包含了全部与通信相关的工作,从而使证认计算部分的各个子任务可以无须通信而独立完成,最大限度地利用了 MapReduce 模型的特性,保证了证认过程的高效性。实验证明,在上亿条的数据量级上此方法可以快速有效地完成交叉证认工作,其性能优于多核环境下基于数据库的并行交叉证认算法以及其他先前的方法,并且随着集群节点个数的增多,其性能表现出了接近于线性的加速比。可见,本文方法的提出为实现实时大规模交叉证认平台打下了基础,在提高天文学家对当今海量天文数据的利用效率方面起到了重要作用。

参考文献:

[1] GRAY J, SZALAY A, BUDAVRI T, *et al*. Cross-matching multiple spatial observations and dealing with missing data, MSR-TR-2006-175[R]. Redmond, WA: Microsoft Research, 2006.

[2] GRAY J, NIETO-SANTISTEBAN M A, SZALAY A S. The zones algorithm for finding points-near-a-point or cross-matching spatial datasets, MSR-TR-2006-52[R]. Redmond, WA: Microsoft Research, 2006.

[3] Report on cross matching catalogues, astroGrid[EB/OL]. (2007)[2008-11-09]. http://wiki.astrogrid.org/pub/Astrogrid/DataFederationandDataMining/cross.htm.

[4] Spatial joins and spatial indexing revisted, astroGrid[EB/OL]. (2007)[2008-11-10]. http://wiki.astrogrid.org/bin/view/Astrogrid/SpatialIndexing.

[5] 高丹,张彦霞,赵永恒. 海量多波段星表数据的交叉证认的实现[J]. 天文研究与技术, 国家天文台台刊, 2005, 2(3):186-193.

[6] 高丹. 海量天文数据融合系统的开发与数据挖掘算法的研究[D]. 北京:中国科学院国家天文台,2008

[7] ZHAO Qing, SUN Ji-zhou, YU Ce, *et al*. A paralleled large-scale astronomical cross-matching function[C]//Proc of the 9th International Conference on Algorithms and Architectures for Parallel Processing. Berlin: Springer, 2009:604-614.

(上接第 3315 页)

[15] 王三民. 模糊推理及态势估计研究[D]. 西安: 西安电子科技大学, 2004.

[16] 李弼程, 邵美珍, 黄洁,等. 模式识别原理与应用[M]. 西安: 西安电子科技大学出版社, 2008: 82-83.

[17] 网络舆情“智库”[EB/OL]. [2009-10-20]. http://www.trs.com.cn/news/gsxw/200910/tt20091020-2565.html.

[18] ATANASSOV K. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20(1):87-96.

[19] ATANASSOV K. More on intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1989, 33(1):37-46.

[20] 雷英杰,王宝树. 直觉模糊逻辑的语义算子研究[J]. 计算机科学, 2004, 31(11):4-6.

[21] 雷英杰,王宝树. 直觉模糊关系及其合成运算[J]. 系统工程理论与实践, 2005, 25(2):113-118.