基于粒度商空间下的话题识别与跟踪研究

毛建景,张君君
(郑州工业应用技术学院 信息工程学院,河南 郑州 451150)

摘 要:文中旨在对自然语言所形成的信息流进行话题识别与跟踪,其目的主要针对网络舆论中出现的新话题进行识别,并实现对已有话题的跟踪研究。基于相容商空间粒度下软聚类算法,实现对话题的识别与跟踪,是舆情分析的关键技术。话题识别与跟踪采用软聚类算法,根据相容关系原理,计算距离函数,使话题呈现一定的层次结构,再利用相容隶属函数实现对边界文本的话题确认,形成注明标注信息的语料。同时,结合基于 Ontology 情感分类法,计算与情感词汇中的语义相似度,统计目标情感词汇的倾向性权重,建立基于粒度商空间下的话题识别与跟踪模型,有效地促进话题倾向性的研究,最终实现对网络舆情话题的识别与跟踪,为相关部门监管网络舆情、掌握舆论方向提供指导。

关键词:相容商空间;粒度;话题识别;舆情
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2019)07-0190-04
doi:10.3969/j.issn.1673-629X.2019.07.038

Research on Topic Recognition and Tracking Based on Granular Quotient Space

MAO Jian-jing, ZHANG Jun-jun
(School of Information Technology, Zhengzhou University of Industrial Technology, Zhengzhou 451150, China)

Abstract: This topic aims to identify and track the information flow formed by natural language. Its purpose is to identify the new topics in the network public opinion and to realize the tracking research of the existing topics. The soft clustering algorithm based on compatible quotient space granularity realizes the recognition and tracking of topics, which is the key technology of public opinion analysis. The topic identification and tracking adopts the soft clustering algorithm. According to the principle of compatible relationship, the distance function is calculated to make the topic present a certain hierarchical structure. Then the compatible membership function is used to confirm the topic of the boundary text, and the corpus indicating the annotation information is formed. At the same time, based on Ontology sentiment classification, the semantic similarity in emotional vocabulary is calculated, the tendency weight of target emotional vocabulary is calculated, and the topic recognition and tracking model based on granular quotient space is established to effectively promote the research of topic orientation. Finally, the identification and tracking of online public opinion topics is realized to provide guidance for relevant departments to supervise network public opinion and master the direction of public opinion.

Key words: compatible quotient space; granularity; topic identification; public opinion

0 引 言

随着网络媒体技术的发展,大多社会热点源于网络舆论,对社会舆情的分析也从传统的机械装置逐渐过渡到网络媒体,网络舆情分析是社会舆情分析的重要途径。

如何挖掘出有效的社会热点或敏感问题,以促进网络监管能力,就成为网络舆情分析的主要研究内容[1]。话题识别作为信息跟踪与处理的主要研究技术,其识别精度和准确判断直接影响识别效度。当前,对话题识别的主要研究方式之一就是聚类分析。常用的聚类分析方法有划分聚类法、密度聚类法、层次聚类法、网格聚类法、模型聚类法等。常用的聚类策略大多采用硬聚类,然而硬聚类过程容易造成话题结构的混乱,降低了边界文本识别度及准确度。

1 国内外研究现状及研究任务

1.1 研究现状

话题识别与跟踪技术作为舆情分析的主要技

收稿日期:2018-08-06 **修回日期**:2018-12-10 **网络出版时间**:2019-03-21
基金项目:河南省高等学校青年骨干教师培养计划项目(2016GGJS-182)
作者简介:毛建景(1980-),女,硕士,讲师,研究方向为计算机网络、现代教育技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0909.036.html

术[2],参与研究的技术人员越来越多,其研究范围也从传统的普通 Web 页面,逐步扩展到微博、博客、新闻 Web、Facebook、论坛等。由于媒体渠道不同,关于话题识别与跟踪技术的研究方法也存在争议。目前,国内外都进行了大量的理论和实践研究。例如,刘倩等对基于情感 Ontology 的资源分析,利用词汇特征抽取的方式对文本的倾向性进行分析[3];史仁仁等提出了周期分类的概念,利用 Single-Pass 聚类算法,完成对网络舆情的分析与研究;周丹晨采用 WordNet,利用上下文本信息同时构造设计出小灵通定位系统(LSC),基于该系统的文本信息描述采用单向路径的聚类算法用以解决对新出现事件的检测问题。

1.2　研究任务

文中以网络话题识别研究为目的,采用软聚类算法,首先计算出距离函数 $\mathrm{dis}(\alpha,\beta)$,并通过与半径 d_i 的比较,在相容商空间粒度下,实现对文档信息的反复分析、连续分解和不断合成,以同步达到聚类重心点集合的形成;其次,利用基于隶属度函数的容度决策理论,即函数 $\mu(X_j,X)=|I(X_j)\cap X|/I(X_j)$,测量出边界文本发生的可能性概率,从而确认具有明确话题标注的信息[4]。通过该方法,可以实现对话题的识别、话题容错、精度确认及话题的跟踪研究,也可有效解决细粒度划分和情感分类等知识共享问题。这些研究在后续文本趋势分析中有重要的理论意义和广阔的应用前景。

1.3　话题识别与跟踪技术概述

(1)概念。话题识别与跟踪(TDT),包括话题识别和跟踪。该技术的出现源于网络信息爆炸下衍生的新问题,目的是解决在线媒体信息流中对话题的识别和跟踪问题[5]。(2)作用。该技术可以识别和跟踪某一特定环境下发生的事情,更能拓展到相关外延事件,从而将话题识别与跟踪的研究领域跨越到对突发事件甚至"未然态"信息的处理。与其相关的定义包括事件(Event)、活动(Activity)、话题(Topic)及报道。

(3)任务。话题识别与跟踪主要完成:对新闻报道的切分,也就是将稿件划分成独立模块;对于第一次出现新的报道的识别;Story Link Detection,即关联性检测,主要目的是检测两篇报道是否属于同一话题;对话题的跟踪,抽取某一特征集以进行匹配为主要任务[6]。

(4)评价。常用的评价形式有评估矩阵,以矩阵形式计算话题的查全率和召回率。

召回率公式为:$R=a/(a+c)$

其中,a 是系统判定属于话题;c 是系统判定不属于话题。

查全率公式为:$F=a/(a+b)$

其中,a 是系统判定属于话题;b 是系统判定不属于话题。

二者之比为调和平均值:$p=2/(1/R+1/F)$

1.4　基于 Ontology 的情感分类体系

基于 Ontology 的情感分类体系是通过词汇语义[7-8],判别其相互之间的相似程度,从而为文本的倾向性提供分析依据。情感 Ontology 中的词汇量的来源有多种途径,其中主要来源于网络数据库,这些数据源是通过相似度计算为理论基础。表达情感的词汇通常只有正面和负面的词汇。对于词汇倾向性的计算,一般需要基于语义相似及情感深度[9]。语义相似度的计算公式为:

$$\mathrm{Sim}=\mu\mathrm{Hsim}(T_i)+(1-\mu)0\mathrm{dis}(T_i)$$

其中,$\mathrm{Hsim}(T_i)$ 是词汇与情感 Ontology 的相似度计算;$0\mathrm{dis}(T_i)$ 是词汇在情感 Ontology 中的深度;μ 为可调节参数,且有 $0\leqslant\mu\leqslant1$。

情感 Ontology 采用向量空间模型来表示文档信息,通常可以将文档表示成:$D_i=\{(T_{i1},w_{i1}),(T_{i2},w_{i2}),\cdots,(T_{in},w_{in})\}$,其中 $T_{ij}(j=1,2,\cdots,n)(T_{i1},w_{i1})$ 指的是文本中的词汇,$w_{ij}(j=1,2,\cdots,n)$ 指的是 T_{ij} 所对应的权重。在进行情感倾向分析时,一般分两步完成:第一,过滤掉不相似的词汇,需要利用相似度公式来完成;第二,对情感倾向性进行判断和识别,该过程要通过分析模型中的权重进行判别。

2　相关技术及基本原理

2.1　相容商空间理论的粒度变换原理

解决问题的过程可以用三元组 (a,b,c) 来表示,其中 a 表示所研究对象的通用名称,也称为论域,函数 b 表示从 a 到 c 的一个映射,c 属于论域的结构,反映 a 中各元素之间相互存在的关系。在对 (a,b,c) 的分析和求解过程中,主要是指对论域 a 及其相关结构和属性的分析以及研究计算。当从不同的粒度进行分析和处理问题时[10],将最细的粒度看作为 a,然后以粗角度分析并以某种方式简化它,对于特征性质相近的作为等价处理。最后,整体作为一个元素构成一个新的域,也是最大的粒度,称之为 $[A]$,并将之前的 (a,b,c) 转化成 $[A,B,C]$。在简化元组的过程中,仿照数学中商集的概念,把不同粒度世界的世界模型称为相容商空间[11]。此时,用 (a,b,c) 对一个问题进行描述,并在其论域上引入等价关系 T,对应于 T 的商集 $[A]$ 作为一个新的论域。在进行分析研究时,对待不同的问题就可以表述成不同的粒度世界,这样就达到了简化问题、解决问题的目的。相容商空间因其强大的表达能力[12],既可以对多种函数进行定义,又可以对论域中的不同元素进行描述,从而分析出不同元素

之间的关系以及结构和运算等。

2.2 相容商和粒度计算的基本简介

粒度计算是一种涵盖所有关于粒度基本理论方法、相关技术及研究工具的新的概念和计算公式。其应用领域主要是分析和处理无法确定和不完整的模糊信息[4],属于软计算科学的一个分支。

相容商空间:假设 (X,Y) 属于拓扑空间,其中 Y 是 X 的拓扑。假设 T 是 X 上的等价关系,则可以计算 X 相应的商集,称为 $[X]$。然后,假设在 $[X]$ 的定义上,将 T 值诱导计算出来,称为 $[T]$。则 $([X],[Y])$ 都是商集的拓扑空间。

假设 R 是相容的,若 $x,y \in (X,Y)$ 并且 $x < y$,则表明 $[x] < [y]$,其中 $[x],[y] \in (X,Y)$。一般情况下,在 (x,y,t) 上的性质,在 $[x]$ 上就不一定具备。如果这些性质不重要,它们在计算分析中自然会被忽略,但这些属性是基于分类抽象的[13],并且一些细粒度信息将相应地消失。因此在实际的计算分析中,要尽量保留这些主要性质。

2.3 相容商空间粒度原理

2.3.1 不同相容商空间粒度的获取

定义 1:令 $[X]=\{x \mid I_x \in X\}$,其中 $[X]$ 是与相容关系 I 相关的相容商空间。根据公式中的关系定义,可以计算出距离函数。假设 a,b,c 都是论域 X 中的三个向量,那么 $\mathrm{dis}(a,b)$ 就是一个距离函数关系。作为距离函数,$\mathrm{dis}(a,b)$ 满足以下特征:

(1) $\mathrm{dis}(a,b) \geqslant 0$;

(2) $\mathrm{dis}(a,a)=0$;

(3) $\mathrm{dis}(a,b)=\mathrm{dis}(b,a)$;

(4) $\mathrm{dis}(a,b) \leqslant \mathrm{dis}(a,c)+\mathrm{dis}(b,c)$。

根据上述条件可以得知,$\mathrm{dis}(a,b) \leqslant d$ 就是一个相容关系,其中也要满足条件 $d \geqslant 0$,也可将 d 称之为函数 $\mathrm{dis}(a,b)$ 的半径。

根据上述条件公式得出,相容关系 I 与距离函数 $\mathrm{dis}(a,b) \leqslant d$ 之间就形成了一种一对一的对应关系。

定义 2:假设 I_1 和 $I_2 \in I$,那么对于任意 $x,y \in X$,都有 $xI_2y \Rightarrow xI_1y$,则称相对 I_2,I_1 更细,表示为 $I_1 < I_2$。

根据定义 2,获得 n 层层次结构对应的 n 个相应的相容关系的序关系:

$$I_0 < I_1 < \cdots < I_n$$

可以通过以上相容序关系及距离函数获得 n 层层次结构,其距离半径有如下序关系:

$$d_0 > d_1 > \cdots > d_n > 0$$

设 I_i 对应的相容商集为 $[X]_i (i=0,1,\cdots,n)$,则不同层次的粒度论域集有如下的相容序关系:

$$[X]_0 < [X]_1 < \cdots < [X]_n$$

根据不同级别层次的粒度论域集的相容序关系,可以得到不同相容商空间的粒度。

定义 3:假设 $\mathrm{IS}=(U,A)$ 是一种信息系统,$X,Y \subseteq A$,则:

(1)若 $x \rightarrow y$,则 $\mathrm{dis}(X) \geqslant \mathrm{dis}(Y)$;

(2)若 $x \leftrightarrow y$,则 $\mathrm{dis}(X)=\mathrm{dis}(Y)$。

由此可以得出,如若 $X,Y \subset A$,则有 $y \rightarrow x$,从而得出 A 属于子集,随着属性的增加,粒度不断减小,则表明分辨率在不断增加。

定义 4:$\forall x \in X$,令 $[x]=\{y \mid (x,y) \in I\}$,称为 $[x]$ 的相容类。

2.3.2 相容商空间粒度下的软聚类原理

(1)在所有数据中,选取最初始的 T 个样本 $Y=\{X_1,X_2,\cdots,X_t\}$ 表示样本的重心点数据的所有集合,同时 d_n 仍旧表示相容空间的距离半径。

(2)通过计算 $\mathrm{dis}(X_a,X_b)$,$a \in (1,2,\cdots,t)$,$b \in (1,2,\cdots,t)$,就能够得出 $\mathrm{dis}(X_a,X_b)$ 和 d_0 之间存在的关系。

(3)通过 $\mathrm{dis}(X_a,X_b) \leqslant d_0$,就可以计算出原来所有样本和重心之间的距离,还能够与距离半径进行比较。

(4)充分利用相容商空间粒度分析法,对通过软聚类计算得到样本重心点之间的距离调整进行反复分析比对[14]。在实际的计算解答中,也可以采用合并求解法对粒度之间的关系进行调整,还能够实现结构层次的划分。

(5)对于边界距离 $\mathrm{dis}(X_a,X_b)=d_a$,则表示 X 值在两个结构中都同时存在,利用任何一个结构公式都可以进行解答计算,通过不断的反复解答计算,就可以得出软聚类的结果。

通过不断的分析和计算,就可以得出距离函数和相容关系之间一对一的对应关系,也可以解决话题层次和不确定边界存在的一些问题。

2.4 相容商空间粒度下的软聚类设计

本节提供了话题识别和跟踪的算法基础。相容商空间粒度的确定是连续不断地分析、比较和调整样本重心点集的过程。在软聚类设计时,通过合并和分解来调整粒度[15],以实现层次结构的明确划分;利用相容隶属函数确定边界。基本路线:(1)选取初始 k 个样本 $X=\{X_1,X_2,\cdots,X_k\}$ 作为样本的重心点集合,并以 d_0 作为相容距离的半径;(2)计算相容距离函数 $\mathrm{dis}(X_i,X_j)$,$i \in (1,2,\cdots,n)$,$j \in (1,2,\cdots,n)$,判断 $\mathrm{dis}(X_x,X_y)$ 与 d_0 的关系;(3)当 $\mathrm{dis}(X_i,X_j) < d_0$ 时,表示 X_i,X_j 属于同一类,同理计算所有样本与重心的距离,并与距离半径进行比较;对于边界距离可表示为 $\mathrm{dis}(X_i,X_j)=d_i$,指的是 X_j 能够同时在两个簇中存在,接着利用相容隶属函数 $\mu(X_j,X)=|I(X_j) \cap X| / I(X_j)$

完成 X_j 所属簇的判断。重复该聚类过程,最终准确地呈现软聚类的结果。

2.5　话题识别与跟踪

(1)文档向量空间降维。解析 Web 语料库中的 XML 文档集并将其表示为向量空间模型 $D=\{D_1,D_2,\cdots,D_r\}$ (D_i 为向量空间, i 为第 i 篇文档)。$D_i=\{(T_{i1},w_{i1}),(T_{i2},w_{i2}),\cdots,(T_{im},w_{im}),\cdots\}$, w_{ij} 表示词汇权重值,指的是文档信息 D_i 中第 j 个词汇的权重。由于某些词汇与话题关联度不高或词频较低,影响话题分析的精度[16],因此需对向量空间降维,抽取与已知话题关联度高的词汇和高频词汇,形成 n 维文档向量空间集 $D=\{D_1,D_2,\cdots,D_r\}$,其中 $D_i=\{(T_{i1},w_{i1}),(T_{i2},w_{i2}),\cdots,(T_{im},w_{im}),\cdots\}$ $(n<m)$ 。

(2)话题层次划分与不确定话题边界确定。采用软聚类算法对目标文档 M 进行识别。经过识别后会形成一个层次话题集,即 $TP=\{tp_1,tp_2,\cdots,tp_s\}$ 。在聚类过程中,动态地形成向量集 $C=\{c_1,c_2,\cdots,c_s\}$,其被称为话题重心点向量集。

(3)话题标题解析。把重心点向量集反馈到预处理的 XML 文档集,解析重心向量集得到标题信息,作为话题标题,形成带有标题、具有层次的话题集 $TP=\{(tp_1,name_1),(tp_2,name_2),\cdots,(tp_s,name_s)\}$ 。根据话题 tp_i 中的所有文档向量集 $tp_i=\{d_{i1},d_{i2},\cdots,d_{ic}\}$ $(0<c\leqslant r)$,更新 XML 文档集中话题节点的标注信息。

(4)新报道向量空间软聚类与话题节点信息更新。根据层次话题集 TP 中的文档向量集 $tp_i=\{m_{i1},m_{i2},\cdots,m_{in}\}$,对 Web 语料文档集中跟话题节点相关的标注信息进行修改更新。确定话题识别后,需要动态跟踪,并在话题监督下完成该过程。

对话题集 $TP=\{(tp_1,name_1),(tp_2,name_2),\cdots,(tp_s,name_s)\}$ 和重心点向量集 $C=\{c_1,c_2,\cdots,c_s\}$,利用软聚类算法对新报道向量空间 V 进行分类。

第一步,计算向量空间 V 与文档重心点集距离函数 $dis(F,C)$;第二步,根据以上距离函数的结果,与相关的距离半径 d_r 进行比较,准确地得出 V 所属的类别。如果获得的距离函数结果大于距离半径,则使用向量空间 V 为重心点,作为新话题加入新层次话题集:$TP=\{(tp_1,name_1),(tp_2,name_2),\cdots,(tp_s,name_s),(F,name_f)\}$,同时,更新 Web 语料库中文本文档话题节点的标注信息。

3　结束语

在相容商空间中,粒度计算可以基于原始的知识来变换和分析各种子集。在以不同层次粒度上的论域、结构和属性对待同一问题进行递进求解时,就可以利用商空间中细粒度和粗粒度之间的保真性定理执行空间层次结构的反复推理和计算,最终得出结果,这种计算方式很大程度上降低和简化了问题在求解过程中的难度。

参考文献:

[1] 李　岩,韩　斌,赵　剑. 基于短文本及情感分析的微博舆情分析[J]. 计算机应用与软件,2013,30(12):240-243.

[2] 冯兆旭. 面向网络舆情分析的社会热点话题技术研究[D]. 兰州:兰州交通大学,2017.

[3] 刘　倩,陶县俊,王晓东. 基于情感 Ontology 的资源分析模型[J]. 计算机与数字工程,2009,37(9):115-119.

[4] 周丹晨. 基于粒计算面向工艺实例检索的材料相似度算法[J]. 机械工程学报,2014,50(13):170-177.

[5] 丁媛媛. 基于时间序列的微博热点话题识别与追踪[D]. 西安:西安科技大学,2017.

[6] 孟　军. 相容粒计算模型及其数据挖掘研究[D]. 大连:大连理工大学,2012.

[7] ANTHONY M X, MARGRET S A. Case-based reasoning (CBR) model for hard machining process[J]. International Journal of Advanced Manufacturing Technology,2012,61(9-12):1269-1275.

[8] 王伦文,张贤骥,张　铃. 基于模糊相容关系的聚类粒度分析[J]. 系统仿真学报,2017,26(7):1492-1496.

[9] 张　铃,张　钹. 模糊相容商空间与模糊子集[J]. 中国科学:信息科学,2011,41(1):1-11.

[10] WANG Lijuan, YANG Xibei, YANG Jingyu, et al. Relationships among generalized rough sets in six converings and pure reflexive neighborhood system[J]. Information Sciences,2012,207:66-78.

[11] 王艳茹,温长峰,洪晓蕾. 相容商空间粒度下的话题识别与跟踪[J]. 中国管理信息化,2011(14):77-78.

[12] 李　霞,王连喜,路美秀,等. 基于复合词生成的网络热点话题识别及描述算法[J]. 图书情报工作,2016,60(23):128-134.

[13] ZHANG X H, DENG Z H, LIU W, et al. Combining rough set and case based reasoning for process conditions selection in camshaft grinding[J]. Journal of Intelligent Manufacturing,2013,24(2):211-224.

[14] 伍育红. 聚类算法综述[J]. 计算机科学,2015,42(z1):491-499.

[15] JEE T, LEE H, LEE Y. Visualization of document retrieval using external cluster relationship[J]. Journal of Information Science and Engineering,2013,29(1):35-48.

[16] MOHD M, CRESTANI F, RUTHVEN I. Construction of topics and clusters in topic detection and tracking tasks[C]//International conference on semantic technology and information retrieval. Putrajaya, Malaysia: IEEE,2011:171-174.