一种新的 Web 中文文本聚类方法研究

叶宇飞[1] 安世全[2] 代 劲[3]

[1](重庆邮电大学计算机科学与技术学院 重庆 400065)

[2](重庆邮电大学移通学院 重庆 400065)

[3](重庆大学计算机科学与技术学院 重庆 400065)

摘 要 传统的文本聚类缺少语义信息,文本的特征向量高维稀疏,忽略了 Web 文本的特殊性。为了解决这些问题,提出一种 Web 中文文本聚类方法。在基于知网(HowNet)的概念空间基础上过滤非名词,分析文本中重要词汇的语义,对标签特征集与正文特征集进行特征集聚类,再利用改进的 TF-IDF 算法选取两个集合中的特征,最终将文本表示为选取的标签特征集与正文特征集的并集,降低了特征的维度,高效地表示了文本。通过实验验证了其有效性。

关键词 Web 文本聚类 特征降维 知网 文本相似度

中图分类号 TP391.1 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2013.12.058

RESEARCH ON A NOVEL WEB CHINESE TEXT CLUSTERING METHOD

Ye Yufei[1] An Shiquan[2] Dai Jin[3]

[1](*College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

[2](*College of Mobile Telecommunication, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

[3](*College of Computer Science and Technology, Chongqing University, Chongqing 400065, China*)

Abstract Traditional text clustering lacks the semantic information, its text eigenvector is high-dimension sparse, and ignores the particularity of the Web text. In order to solve these problems, we propose a Web Chinese text clustering method in this paper. On the basis HowNet-base concept space, the method filters the terms but nouns, analyses the semantics of the important words in the text, and carry out the feature set clustering on label feature set and text feature set. Then it uses the improved TF-IDF algorithm to select features from these two sets, and finally expresses the text as a union of the selected label feature set and text feature set. It reduces the dimensions of features, and expresses the text efficiently. Experimental results demonstrate its effectiveness.

Keywords Web text clustering Feature dimension reduction HowNet Text similarity

0 引 言

文本聚类是一个将文本集划分为若干组或类的过程,并使得同一组内的文本对象具有较高的相似度,而不同组中的文本对象则差别较大。对于 Web 文本聚类研究比较早的有 zamir 等人提出的聚类器是第一个基于 Web 搜索结果的聚类系统,它采用后缀树聚类算法,Web 搜索结果被表示成树结点。把具有相同词和词组的 Web 搜索结果汇成簇,用出现最频繁的特征词作为类标记。

1 相关研究现状

按照聚类分析算法的主要思路,聚类可分为划分法、层次法、基于密度法、基于网格法、基于模型法和基于概念的方法。划分法最常用的有 K-Means 和 K-Medoids 算法,缺点是对于 K 值的不确定性。层次法的主要缺陷是不能回溯,即不能更正错误的决定。密度法虽然克服了"类圆形"聚类的缺点但计算密度单元的复杂度大,可能会造成频繁的 I/O 操作。网格法的代表算法有:STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法,这类算法采用量化空间,因此处理速度较快。基于模型法尝试着寻求给定数据与一些数学模型的最佳匹配,大致可以分为基于统计的方法和基于神经网格的方法,包括关联规则法、决策树、竞争学习、SOM 算法等。

基于概念的聚类方法是从语义的角度分析文本的构成,典型的概念聚类方法有 LINGO,它首先从 Web 文本搜索结果确定概念文档,并抽取特征词作为类标记,把与类标记相似的文档聚合成类。德国卡尔斯鲁厄大学的 Hotho 和 Staab 提出了基于本体论的文本聚类,将 Wordnet 作为背景知识来构造文本特征空间,以及 Bhogal J 提出的基于本体的文本检索模型,开创了将本体论应用于文本聚类的先河。印度理工学院的 Bhoopesh Choudhary 和 Pushpak 提出了基于语义聚类的方法[1]。

这几种基于概念的文本聚类算法在实验中都取得了良好的

收稿日期:2012-08-16。叶宇飞,硕士生,主研领域:Web 文本控制,人工智能。安世全,教授。代劲,博士生。

结果。但也存在一些问题:很多聚类方法大多是用空间向量模型来表示 Web 文本的,文本特征词的维数过高,从而使计算复杂度增加。而现有的概念格、潜在语义标引等方法虽然在一定程度上解决了特征维度的问题,但实际上仍然是基于统计的聚类方法,由于没有考虑特征的语义信息,影响了文本聚类的准确性[2]。例如一篇关于篮球和一篇关于足球的两篇文章都描述了体育运动,应该说具有很强的相关性。但在向量空间模型中,可能由于两者的特征词不同,而认为是不相似的两篇文章。相反,一篇讲苹果(水果)和一篇讲苹果(手机),可能被聚类到同一类别,这样最终的聚类结果也就与人们的直观感受相去甚远。另外,许多关于文本聚类的研究并没有注意到普通文本聚类与 Web 文本聚类的区别与联系,使文章中的方法很难直接应用在 Web 中。

本文针对上述方法存在的问题并结合 Web 文本半结构化的特点,提出了一种既能有效降低文本表示模型的维度,又能结合词项语义信息进行相似度量计算的 Web 中文文本聚类方法。该方法首先采用了词性过滤进行文本预处理,删减了大部分无用的或对区分文本贡献不大的特征词,其次通过特征语义聚类使得文本特征集更能体现文本的语义内涵,再利用基于 Web 文本半结构化而改进的 TF-IDF 算法对文档的特征集进行筛选,经过以上三层处理之后,得到本文中高效低维的特征集,建立了一种基于知网语义特征并体现特征分类强度的文本表示模型,以特征词间的语义相似度衡量文本间的相似度。通过实验分析,验证了本文提出的聚类方法的有效性。

2 知　网

2.1 知网结构简介

《知网》是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[3]。《知网》中有两个主要的概念:“概念”与“义原”[4]。“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的,这种“知识表示语言”所用的“词汇”叫作“义原”。“义原”是用于描述一个“概念”的最小意义单位。

在《知网》中,一共描述了义原之间的 8 种关系:上下位关系、同义关系、反义关系、对义关系、属性与宿主关系、部件与整体关系、材料与成品关系、事件与角色关系。根据义原的上下位关系,所有的“基本义原”组成了一个义原层次体系。

2.2 基于知网的语义相似度计算方法

因为所有的概念都最终归结于用义原(个别地方用具体词)来表示,所以义原的相似度计算是概念相似度计算的基础。所有的义原根据上下位关系构成了一个树状的义原层次体系,所以计算语义距离可采用如下公式:

$$Sim(p_1,p_2) = \frac{a}{d+a} \tag{1}$$

其中 P_1 和 P_2 表示两个义原,d 是 P_1 和 P_2 在义原层次体系中的路径长度,是一个正整数。a 是一个可调节的参数。

概念的相似度计算是以义原的相似度计算为基础的,一个概念由多个义原描述,这些义原描述的方面各不相同,不论义原属于哪种关系,本文把所有义原都作为同种义原对待,保持其原有的顺序后,每个概念对应一个义原集合,参考文献[5]的方法,使用以下公式计算概念的相似度:

$$Sim(W_1,W_2) = \frac{Same + Sim(i,j)}{Same + 1} \tag{2}$$

其中,$Same$ 为词语 W_1 和 W_2 的义原描述集合中相同义原的数目,$Sim(i,j)$ 是 W_1 和 W_2 的义原描述集合中第一对不相同的义原的相似度。

3 Web 中文文本聚类

传统文本聚类算法并不能直接应用到 Web 文本聚类上。一方面,尽管 Web 文本挖掘和通常的平面文本挖掘有类似之处,但是许多文本聚类算法的实现都依赖于数量有限的离线文档集。而 Web 文本聚类的对象一般是网络搜索引擎返回的查询结果,或者具有特定主题的 Web 文档集合,这些文档集数庞大,并且具有高流动性,传统的文本聚类算法无法满足在线计算对于效率、质量的要求。因此,必须根据 Web 文本挖掘的特点选择合适的聚类算法。Web 文本聚类算法需满足的关键条件为以下六个方面:相关性、产生可浏览的“簇”信息、交叠性、摘要容错性、速度、增量性[6]。从实际应用背景出发,这六个条件限定了文本聚类算法的选择空间,并为 Web 文本聚类算法的改进和创新提供了明确的思路。

3.1 Web 文本获取

Web 文本获取的方式主要是通过广为人知的网络蜘蛛或者网络爬虫,即 Web 文本采集器抓取的。Web 文本采集器能够自动漫游于 Internet 站点之间,按照某种人为设计的策略或者算法进行远程数据的搜索与获取,并将获取的文本存储在系统的本地文本库中,用于进一步的分析利用[7]。

3.2 Web 文本预处理

Web 页面上包含着许多跟文本聚类无关的内容,除了前文提到的 < TITLE > 、< KEYWORDS > 、< DESCRIPTION > 这些有意义以外的 HTML 标签,以及脚本程序等都称作“噪音”,它们影响着文本挖掘的准确率和效率。因此必须进行本文预处理,去掉跟文本挖掘无关的“噪音”。在去掉各种标签的时候,由于上述三种标签中的内容对于表征文本类别有重要的意义,所以需要特殊的处理:将其保存到单独的文本文件中,而文本的正文则保存在另一文本内。去除掉这些噪音后,每一篇 Web 文本都由两部分组成:标题、关键字、描述标签中的纯文本;正文组成的纯文本。

由于虚词如感叹词、介词、连词等对于标识文本的类别没有意义,而表征文本特性的往往是文本中的实词,如名词、动词、形容词、副词等,其中又以名词的信息量相对最大。因此,在对每篇文本的两部分进行分词时,过滤掉所有的非名词,分词后的名词集合文本依然分作两部分,即 < TITLE > 、< KEYWORDS > 、< DESCRIPTION > 标记中的名词以及正文中的名词。

定义 1　标签特征集:经过预处理和分词后,由 < TITLE > 、< KEYWORDS > 、< DESCRIPTION > 标签中的名词组成的特征集合。

定义 2　正文特征集:经过预处理和分词后,由正文中的名词组成的特征集合。

3.3 文本特征提取

3.3.1 特征排歧

由于某些词语可以描述为多个概念,各个概念的含义并不

相同甚至相差很大,因此需要对标签特征集以及正文特征集中的名词选择合适的概念定义,本文采用选择最大语义场密度[8]的方法进行语义排歧。

3.3.2 特征集语义聚类

经过大量的数据统计表明,一个文本的主题可以被若干个彼此之间语义相似度较大的特征词表示,并且标签特征集中的某些名词必定跟主题非常相关。根据前文中提到的基于知网的概念词语、义原的相似度计算方法,进行语义特征聚类,具体算法描述如下:

(1) $D = \cup_{i=1}^{k} W_i$ 且 $D = D_1 \cap D_2$,文本 D 经过分词和预处理,共有 K 个词语 $W_1, W_2, \cdots, W_j, W_{j+1}, \cdots, W_k$,分为 D_1、D_2 两个集合,其中前 j 个词语是标签特征集为 D_1,而后 $k-j$ 个是正文特征集即 D_2,它们仍然分别保存在单独的文本中。

(2) 找到 D 的一个真子集 d,用以表征该文本,具体做法如下:

a) 初始化 j 个集合,$d_i = \{W_i\}, W_i \in D_1, i \leqslant j$。

b) 根据 1.2 节的方法计算 D_1 中所有词语的相似度 $Sim(W_m, W_n), W_m, W_n \in D_1$. 一共有 $j \times (j-1)/2$ 个结果,将这些结果按降序排列。

c) 依次取出特征项相似度,选择相似度大于阈值 a 的两个词语归为一类,将它们所在的集合合并。

d) 选择最后剩余的集合中,拥有元素最多的一个集合记为 d_1。

e) 对 D_2 重复上述步骤,得到集合 d_2。

f) $d = d_1 \cap d_2$。

3.3.3 特征选择

TF-IDF 方法是文本相似度量的方法中最为典型的一种。该方法基于词频与逆文本频率的经验观察,将文本表示为文中出现的 n 个加权词项组成的向量[9]。计算词项 W_i 的 TF-IDF 权重通常采用如下公式:

$$TFIDF(W_i) = TF(W_i) \times \log\left(\frac{N}{DF(W_i)}\right) \tag{3}$$

其中 N 是所有文本的数目,$DF(W_i)$ 是包含词语 W_i 的文本数目。$TF(W_i)$ 表示 W_i 在当前文本中出现的频率。本文中的当前文本包括标签文本和正文文本。

由于标签特征集中的词语对于表示文本有非常重要的意义,应该加大其在特征集中的权重,采用如下改进的计算公式:

$$TFIDF(W_i) = TF(W_i) \times \log\left(\frac{N}{DF(W_i)}\right) \times Pw_i \tag{4}$$

Pw_i 是新加入的权重因子,根据当前词项 W_i 类属标签特征集或是正文特征集,给予不同的量值,通过这个权重因子来平衡特征项权重对特征项频率的过分依赖。Pw_i 的数值没有固定的取值方法,可先假定一个数值,然后根据大量的实验来进行调整。如果 W_i 同时出现在 D_1、D_2 两个特征集中,则 Pw_i 取标签特征集的权重因子值。本文经过多次实验,确定权重因子的取值规则如表 1 所示。

表 1 权重因子取值表

特征项在 Web 文本中的位置	标签特征集	正文特征集
权重因子 Pw_i	1.7	1

计算所有的 $TFIDF(W_i)$ 以后,按降序排列,选取前 15 个作为最终的语义名词特征集,得到文本的表示模型:$D_i = \{W_{i1}, W_{i2}, \cdots, W_{ik}\}$。其中,$D_i$ 代表文本集合中的第 i 个文本,$i \in [1, N]$,W_{ij} 表示文本 D_i 的第 j 个语义名词特征项,$j \in [1, k]$。

3.4 文本间的相似度度量

聚类的一个重要环节就是找到一个精确的相似度度量方法,聚类算法要以此作为文本对象之间比较的依据,然后对文本聚类。

本文利用基于文本语义聚类的聚类方法,采用如下的文本相似度计算公式:

$$Sim(D_i, D_j) = \frac{1}{m \times n} \sum_{p} \sum_{q} Sim(W_{ip}, W_{jq}) \tag{5}$$

其中 $Sim(D_i, D_j)$ 表示文本 D_i 与文本 D_j 的相似度,m、n 分别为文本 D_i、文本 D_j 的维度,$p \in [1, m]$,$q \in [1, n]$,$Sim(W_{ip}, W_{jq})$ 为特征 W_{ip} 与 W_{jq} 的语义相似度。

3.5 簇特征提取

将多个彼此间的相似度达到阈值的文本划分在一个类别中,这样的文本集合类别称为一个文本簇。而簇特征表示了某个簇,也代表了该簇的所有文本的共同主题。如果能够得到某个簇的簇特征,在实时系统中就可以很方便地计算新获取的数据所属的类别,这对于 Internet 上异构、海量、分布的 Web 文本聚类,具有很高的实用性。

对于某个簇 C_i 包含 K 个文本,每个文本描述为 $D_i = \{W_{i1}, W_{i2}, \cdots, W_{im}\}$,簇特征的提取方法如下:

(1) 将 C_i 中所有文本的特征取出,合并为一个新的特征集:$D_i' = \bigcup_{j=1}^{K} D_{ij}$。

(2) 计算 D_i' 中所有特征项的词频 TF,合并相同的特征项得到:$D_i' = \{(W_1, TF_1), (W_2, TF_2), \cdots, (W_n, TF_n)\}$。

(3) 对 D_i' 中的所有特征词进行特征语义聚类,找到相似度达到阈值的特征词所组成的最大真子集 d_i'。对于所有的标签特征集也视为全部正文特征集的一部分,不再分别计算。

(4) 将 d_i' 中的所有词语按照词频 TF 降序排列,选取前 15 个词语作为该文本簇的簇特征,得到簇表示模型:$C_i = \{W_{i1}, W_{i2}, \cdots, W_{ik}\}$。

由于与文本表示模型形式相同,计算文本与簇、簇与簇间的相似度与计算文本间的相似度方法一样。

4 Web 文本聚类算法

在文本聚类应用中,需要将 n 个不同的文本分在一组不相交的集合中,每一个集合即是一个簇。将每个文本组成一个集合,一共有 n 个集合,而每个集合的簇特征初始化为该簇中文本的特征。将文本集合中任意两个文本 D_i, D_j 的相似度记为 $Sim(D_i, D_j)$。计算所有文本的两两相似度并按降序排列,共有 $n \times (n-1)/2$ 个结果。依次取出相似度值 $Sim(D_i, D_j)$,分别找出 D_i 和 D_j 所属的集合 C_i 和 C_j,将其合并为一个新的集合,并删除原来的集合 C_i 和 C_j。若 $Sim(D_i, D_j)$ 小于阈值,则更新所有簇的簇特征,更新完成后本次迭代结束。重复以上步骤,直到两次迭代后,簇的数目不变,则算法终止。具体算法描述如下:

输入:经过预处理的 N 个未标注的 Web 文本的特征集:$D_1, D_2, \cdots, D_N$,阈值 μ

输出:已标注的 K 个类簇:$C_1, C_2, \cdots, C_K$

(1) 初始化每个簇:$C_i = \{D_i\}, i \in [1, N]$,将文本特征作为簇特征。

(2) 计算任意两个簇的相似度 $Sim(D_i, D_j)$(D_i, D_j 分别是簇 C_i, C_j 的簇特征),按降序排列,得到相似度集合 S。

(3) 顺序取出集合 S 中的相似度值 $Sim(D_i,D_j)$,将 D_i 所在的簇 C_i 与 D_j 所在的簇 C_j 合并为一个簇,删除簇 C_i,C_j。

(4) 若 $Sim(D_i,D_j) < \mu$,本次迭代结束,更新所有簇的簇特征。

回到步骤(2)重新迭代,如果两次迭代结束后,簇的数目不变,则停止迭代。

5 实验与分析

实验 1 *验证聚类方法的有效性*

为了验证本文提出的基于知网和文本结构的 Web 中文文本聚类方法的有效性,使用 LoalaSam 文本采集器从新浪网上抓取了 2 000 篇文档,共分为军事、社会、体育、教育等 4 个类别,每个类别选择 500 篇文档作为实验数据。本文采用查准率和查全率以及 F-度量值[10]对聚类实验结果进行分析,首先比较了阈值 μ 对聚类结果的影响,然后将本文的方法与基于 VSM 的 K-Means 方法以及文献[5]基于知网语义聚类的方法进行对比。

图 1 给出了在选取前 15 个关键词项作为文本特征向量,利用本文的聚类算法进行聚类的条件下,同一聚类中的相似度阈值 μ 的不同对聚类结果的影响。随着 μ 的逐渐升高,聚类效果也逐步提升。这是因为随着 μ 的提高,文本之间的区分度越来越大,使得聚类效果越来越好。但当 μ 在 0.7 ~ 0.75 之间到达聚类的最好效果时,再提高 μ 反而降低了聚类效果。F-度量值在 μ 超过 0.75 后迅速下降,这是因为本文选取的文本相似度算法对于相似度的计算结果很少能够超过 0.75,使得本应该在同一簇中的文本聚类在了多个簇中,因而在整体上降低了 F-度量值。而社会分类的结果比较特殊,F-度量值相对于其他分类普遍较低,而且随着 μ 的增加迅速下降,经过分析发现,原因是该类文本的词汇比较分散,在知网中这些特征项彼此相似度都比较低,不能很好地代表文本的核心内容,这也是后文社会类别的文本聚类效果不佳的直接原因。

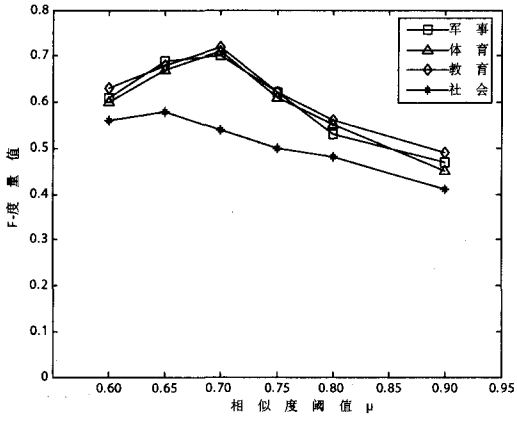

图 1　阈值 μ 对聚类结果的影响

在设定相似度阈值 μ 为 0.7 的情况下,对三种方法进行了对比,结果如表 2 所示。

表 2　聚类结果比较

类别	标准%	军事	社会	体育	教育
基于 VSM 的 K-Means 的聚类方法	查全率	70.5	75.3	71.9	69.8
	查准率	81.6	78.5	80.4	79.1
文献[5]基于知网语义的聚类方法	查全率	85.2	68.4	82.9	80.7
	查准率	90.5	71.3	82.8	87.7

续表 2

类别	标准%	军事	社会	体育	教育
基于知网和文本结构的聚类方法	查全率	89.4	70.5	85.6	84.9
	查准率	91.8	73.7	87.3	88.2

如表 3 所示,本文所使用的方法中每篇文本的特征维度都不会超过 15,因此也不会出现稀疏的向量,降维效果很明显。

表 3　特征维度比较

特征空间维度 / 类别	基于关键词的向量空间	基于知网的概念空间
军事	22 023	6 618
社会	23 231	7 500
体育	19 108	5 289
教育	21 357	6 305

从以上结果可以看出,本文提出 Web 中文文本聚类方法与基于 VSM 的 K-Means 方法相比有较大的优势,而后者几乎达到了瓶颈,很难再有提高。与文献[5]提出的基于知网语义相似度的中文文本聚类方法相比,由于结合了 Web 文本半结构的特点,将标签特征集进行了特别的处理,提高了聚类的查准率和查全率。

实验 2　*对未知类别数据集的聚类结果分析*

在实验 1 中验证了本文方法的有效性的基础上,实验 2 又随机抽取了 420 篇来自新浪网站在 9 月份的新闻。通过聚类处理,结果如表 4 所示,其中阈值 μ 设置为 0.7。

表 4　文本聚类结果

类别	数量	代表词
1	35	篮球 训练 季前赛 常规赛 伤病 恢复
2	46	中国 日本 钓鱼岛 海域 购买 愤怒
3	7	榜 产 贸易 人权 动态 欧盟
4	45	教育 考研 学习 大学 教材 方法
5	29	地产 新房 开盘 优惠 楼盘 贷款
6	27	警察 银行 抢劫 案件 破获 击毙
7	24	中国 好 声音 上海 决赛 选手
8	43	义利观 多哥队 瑞士队 商务部 内地
9	11	镶嵌 周末 宝石 珠宝 黄金 框架
10	31	力作 笔记本 寻址 展会 逻辑 处理器
11	46	演戏 美国 美军 航母 太平洋 军舰
12	76	国庆节 中秋节 长假 旅游 游客 交通

从表 4 可以看到,聚类结果分为 12 个类别,每个类选取其簇特征的前 6 个代表矢量表示。同时可以发现每个聚类代表点都具有明显的意义,如第 12 个聚类代表点数量为 76,数量最多,代表词为"国庆节 中秋节 长假 旅游 游客 交通",由此,可以估计该聚类为近期的热点,主要包括了国庆节中秋节长假出行旅游的相关问题的新闻。

6 结　语

本文较为详细地阐述了 Web 文本聚类过程中涉及到的相关技术。与传统方法不同,本文方法通过改进的 TF-IDF 算法进

(下转第 287 页)

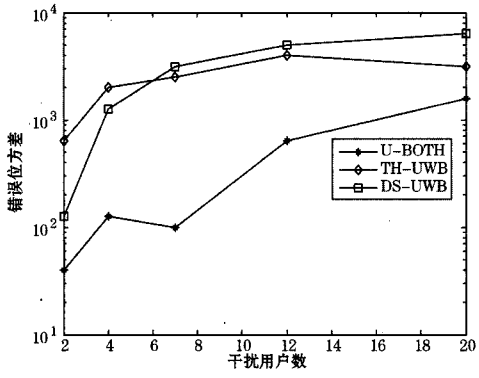

图 3　错误位方差随着用户数目的变化情况

为了评价测距算法的性能,与 Cramer-Rao(CRLB)方法比较了参数估计误差。

图 4 说明 d 和 CRLB 间的关系,以及 d 和均方误差(MSE)的关系。一方面,CRLB 和 MSE 随着 d 的增加而增加,另一方面,测距的 MSE 和 CRLB 比较接近。当 d 非常小时,它们甚至彼此重叠。迭代次数越多,CRLB 和 MSE 的差别就越小。当 $N=20, d>20$ m 时,测距估计的 MSE 上升超过 1 m^2。因此,需要过滤掉较大的 d 以达到更高的测距精度。

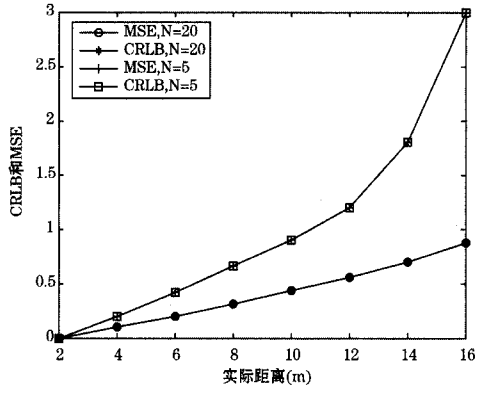

图 4　不同距离下的 CRLB 与 MSE 对比

5　结　语

本文提出一种井下无线传感器网络中使用 UWB 通信技术的快速定位方法。UWB 通信包括一种称为 U-BOTH 的新型 UWB 编码方法,以及类似于 ALOHA 的信道接入方法和采集定位信息的消息交换方法。在此基础上,提出一种测距定位算法,采用极大似然估计计算距离路径衰减模型下的节点距离,并使用最小二乘法计算坐标。最后,使用仿真模拟对 U-BOTH 通信系统和定位算法进行了性能分析。结果表明,U-BOTH 技术能够有效地减少路径损耗模型下的误码率,测距和定位算法能够准确定位井下移动目标。

参　考　文　献

[1] Wang N, Zhu Y, Wei W, et al. One-to-Multipoint laser remote power supply system for wireless sensor networks[J]. IEEE Sensors Journal, 2012,12(2):389-396.

[2] 冯延蓬,仵博,郑红燕. 基于 FPOMDP 的无线传感器网络动态调度算法[J]. 计算机应用与软件,2012,29(8):55-58.

[3] Wang F, Liu J C. Networked wireless sensor data collection: issues, challenges, and approaches[J]. IEEE Communications Surveys & Tutorials,2011,13(4): 673-687.

[4] Sun K, Ning P, Wang C. Secure and resilient clock synchronization in wireless sensor networks[J]. IEEE Journal on Selected Areas in Communications ,2006,24(2):395-408.

[5] Sakairndr P, Ansari N. Security services in group communications over wireless infrastructure, mobile ad hoc, and wireless sensor networks [J]. IEEE Wireless Communications ,2007,14(5): 8-20.

[6] 谢晓松,程良伦. 传感器网络基于移动信标改进的 DV-Hop 定位算法[J]. 计算机应用与软件,2011,28(4): 84-87.

[7] Molisch A F, Cassioli D, Chong C C, et al. A comprehensive standardized model for ultrawideband propagation channels[J]. IEEE Transactions on Antennas and Propagation,2006,54(11):3151-3166.

[8] Win M Z, Scholtz R A. Ultra-Wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communication [J]. IEEE Transactions on Communication,2000,48(4) 679-691.

[9] Yuan J, Wei Y. Joint source coding, routing and power allocation in wireless sensor networks[J]. IEEE Transactions on Communications , 2008,56(6):886-896.

(上接第 225 页)

行特征选择,受到文本聚类的启发,利用特征语义聚类将文本表示为一组彼此语义相似度高的名词模型,同时考虑了具有重要意义的标签文本中的特征,最终将文本表示成筛选后的标签特征集与正文特征集的并集,高效精炼地表达了 Web 文本。

在未来的研究工作中,将针对知网相似度计算的有关缺陷,对计算方法进行完善,更精确地计算词汇之间的相似度。同时对于本文方法处理主题较分散的文本聚类效果不佳的情况,改善本文的文本聚类方法。另外,在知网这个常识知识库的基础上,组建领域知识库,通过构建领域特征集为领域文本聚类提供依据和平台也是今后的研究方向。

参　考　文　献

[1] 李云,田素方. 基于概念格的文本聚类[J]. 计算机工程与应用,2008,44(23):169-172.

[2] 彭京,杨冬青,唐世渭,等. 一种基于语义内积空间模型的文本聚类算法[J]. 计算机学报,2007,30(8):1354-1363.

[3] 董振东,董强. "知网"[OL]. http://www.hownet.com.

[4] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 计算机语言学及中文信息处理,2002,7(2):59-76.

[5] 许君宁. 基于知网语义相似度的中文文本聚类方法研究[D]. 西安电子科技大学,2010.

[6] Oren Zamir, Oren Etzioni. Web Document Clustring: A Feasibility Demonstration[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development In Information Retrieval,1998:46-54.

[7] 许高建. 基于 Web 的文本挖掘技术研究[J]. 微机发展,2007,17(16):187-190.

[8] 白秋产,金春霞,周海岩. 概念向量文本聚类算法[J]. 计算机工程与应用,2011,35(35):155-157.

[9] Salton G. The SMART Retrieval System Experiments in Automatic Document Processing Englewood Cliffs[C]. New Jersey: Prentice Hall Inc, 1971.

[10] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 的文本相似度量方法[J]. 计算机学报,2011,34(5):856-864.