基于词类和搭配的微博舆情文本聚类方法研究

王恒静[1],曹存根[2],高　尚[1]

(1.江苏科技大学计算机科学与工程学院,江苏 镇江 212003)
(2.中国科学院计算技术研究所智能信息处理重点实验室,北京 100190)

[**摘要**]　微博是近年出现的新型互联网信息交流平台,它具有主题分散、体量短小、文体自由等特性,它能对社会产生巨大的影响,所以信息监管部门和商业企业对基于微博信息的舆情分析都有迫切需求. 提出了基于搭配的文本聚类新方法,该方法先进行微博文本预处理,然后利用词类模型进行自动抽取有效搭配,最后基于有效搭配的模型进行文本聚类. 实验证明利用词类文本聚类方法比传统文本聚类方法性能提高 6.3%,而本文方法比利用词类文本聚类方法性能提升了 16.8%,结果显示了本方法的有效性.

[**关键词**]　微博舆情分析,词义类簇,搭配,相似度,文本聚类
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2015)01-0057-09

Research on Text Clustering of Micro-Blog Public Opinion: Word Sense Cluster and Collocation-Based Method

Wang Hengjing[1], Cao Cungen[2], Gao Shang[1]

(1.School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)
(2.Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Micro-blog is the new internet information exchange platform emerged recently, which has the features of theme dispersion, short volume, stylistic freedom, and it can have a huge impact on society. So the information supervision department and commercial enterprise have urgent demand for public opinion analysis based on micro-blog information. This paper presents a novel collocation-based method for text clustering. This method conducts micro-blog text preprocessing firstly, and then uses word sense clustering model to extract effective collocation automatically, and effective collocation-based text clustering finally. Experiments proved that the efficiency of the text clustering method using word sense cluster is higher than traditional text clustering method by 6.3%, and the method of this paper has higher rate than the text clustering method using word sense cluster by 16.8%. The result shows the validity of our method.

Key words: micro-blog public opinion analysis, word sense cluster, collocation, similarity, text clustering

舆情是“舆论情况”的简称,是指在一定的社会空间内,围绕中介性社会事件的发生、发展和变化,作为主体的民众对作为客体的社会管理者及其政治取向产生和持有的社会政治态度. 它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和. 网络舆情是社会舆情在互联网空间的映射,是社会舆情的直接反映[1]. 微博,即微博客(Micro-blog)的简称,是一个基于用户关系信息分享、传播以及获取的平台,用户可以通过 WEB、WAP 等各种客户端组建个人社区,以 140 字左右的文字更新信息,并实现即时分享. 2013 年上半年,新浪微博注册用户达到 5.36 亿,2012 年第三季度腾讯微博注册用户达到 5.07 亿,微博成为中国网民上网的主要活动之一. 微博快速发展及其在引导舆论中的重要作用,需要我们对微博进行舆情监控. 舆情分析建立了百姓和政府、消费者和企业的桥梁[2]. 因此,网络内容监管任务的意义重大但也尤为艰巨,为了建立安全、绿色、和谐的网络环境,必须研究微博的舆情分析技术. 由于微博的信息量巨大并且在互联网上广泛传播,给准确有效地进行微博舆情监管带来了一定的困难.

收稿日期:2014-08-16.
基金项目:人工智能四川省重点实验室开放基金(2012RYJ04)、中科院智能信息处理重点实验室开放课题(IIP2013-1).
通讯联系人:王恒静,研究生,研究方向:自然语言处理. E-mail:hjwang200810@163.com

近年来,不少专家在研究更准确的方法. 对文本中的同义词和近义词难以解决的问题,文献[3]提出了基于词义类簇的文本聚类(SCM)方法,旨在处理同义词和近义词测问题. 它与潜在狄利克雷分布[4](Latent Dirichlet Allocation,LDA)相比,在细粒度的信息区分上更有优势. LDA 主题模型是一种非监督机器学习技术,可以用来识别大规模文档集(document collection)或语料库(corpus)中潜藏的主题信息. 它采用了词袋(bag of words)的方法,这种方法将每一篇文档视为一个词频向量,从而将文本信息转化为易于建模的数字信息.

本文提出的基于词类和搭配的文本聚类方法可以解决大量微博短文本的聚类问题并且准确率有所提高,从而更有利于对微博舆情的监管. 该方法首先要构造词义类簇空间,以下把词义类簇简称为词类. 然后寻找词类间的有效搭配,用该搭配来表示微博文本,就可以匹配不止 1 条与主题相关的微博文本. 构造词类有两部分组成,首先利用 LDA 主题模型从语料库中归纳主题,确定主题中的词语,这种词语被称之为词条,该过程称为词义归纳(WSI);然后通过聚类方法合并相同或者相似的词条生成词类,该过程称为词义聚类(WSC);接着把聚类后生成的词类进行有效搭配自动抽取,形成搭配空间,在该搭配空间内表示微博文本;最后对经过有效搭配处理后的微博文本进行聚类. 该方法可以解决微博中的有相同或者相近意思的文本. 1)用这些词类可以解决微博文本中同义词或者近义词的问题,因为同义词或者近义词被聚类到了同 1 个词类中. 2)利用 LDA 主题模型[5]可以把文本中的每个关键词根据它的上下文给予 1 个特定的词类,即使该词条有不同的意义也会被识别到不同的词类中,可以得到更加准确的文本相似度,从而使计算机理解自然语言更加准确.

本文结构安排如下:第 2 节主要介绍了构造词类模型的相关知识;第 3 节主要介绍了自动抽取有效搭配的相关研究;第 4 节详细介绍基于搭配的微博文本聚类的设计过程;第 5 节通过实验验证该方法的有效性和其存在的缺点;最后一节总结全文.

1 词类模型

1.1 词义和词义类簇

定义 1 词义[3,7,8]:特定词 w 的词义 s_w 可以表示为一组上下文的词的概率分布. 如式(1)所示:

$$s_w=\{t_i:p(t_i|s_w)\},\quad i=1,\cdots,N, \tag{1}$$

其中 t_i 表示上下文中的词,$p(t_i|s_w)$ 表示对 t_i 于词义 s_w 的概率.

定义 2 词义类簇:词义类簇指一组由词义聚类算法得到的词义,它可以表示为式(2):

$$c=\{s_w^j\},\quad j=1,\cdots,M, \tag{2}$$

其中 s_w 表示一个词义.

关于词义和词义类簇的例子如下:

例 1 词“骚扰”的词义“骚扰#”

打扰: 0.043

骚扰: 0.138

例 2 词义类簇

{作用#,功能#}

作用#={作用:0.149,功能:0.059,功效:0.029}

功能#={功能:0.169,作用:0.069,功效:0.039}

该词义类簇中,“作用#”和“功能#”上下文概率比较相似,所以被聚到一个类中,形成一个的含有若干词条的词类. 可文本表示为:

作用近类={作用,功能}.

1.2 词类模型

词类模型可以用词类来代替文本中的词条,将微博文本表示在词类空间上,例如有一条关于保险行业的微博文本:

太平洋保险很好.

其中“太平洋保险”在“太平洋保险近类”中,“好”在“好近类”中,用词类替换词语后形成分词后的文

本为:

太平洋保险近类　很　好近类

采用词类来表示文本,该模型不仅解决了多义词还处理了文本中的同义词现象,如例 2 中的词“作用”与“功能”被聚到同一个词类中.

2　自动抽取文本中有效搭配

2.1　搭配的定义

随着计算机技术的快速发展和应用,自动抽取搭配[9]成为越来越多的人重视的自然语言处理任务之一. 但不同的研究者对搭配的概念有不同的理解,现在还不能准确地定义搭配,不同的研究者界定搭配的范围和方法都有所差异. 本章主要是寻找经过基本处理后的保险行业的微博文本中的有效搭配,并介绍自动抽取搭配的方法.

不同的研究者对搭配有不同的定义,文献[10]中提到 Firth 学派采用不同的量化指标来刻画搭配的显著性,他们主张的是属于基于语料库的定量研究方法. 新弗斯学派(neo-Firthians)的代表人物 Halliday 和 Sinclair 发展了 Firth 的理论. 他们提出了搭配的整套概念和方法,如节点词、跨距、搭配词,从大量的语料中提取搭配. 他们认为搭配词和节点词有某种相互吸引力,并用某种方法来测量他们之间的关系,从而确定搭配关系. Firth 学派主张的搭配界定是基于统计的,他更加重视的是搭配词语在语言例子中所体现的统计特征[10,12].

文献[11]中可以得到以 Cowie 为代表的学者从搭配组成词语的特征上定性的定义搭配,是一种“短语法研究”. 他将词语的组合分为 4 类:1)纯习语(Pure idiom);2)比喻性习语(Figurative idioms):词语之间的组合有比喻意义,但字面意义仍在使用,如:catch cold;3)受限搭配(Restricted collocation):一个词仅在该组合中有比喻意义,另一个词保留字面意义并可替换为其他的词;4)开放搭配(Open collocation):两个构成元素可以自由重新组合. 无论是哪种搭配,在和其他组合区分时都要考虑搭配词语之间的意义关系. 也可以看出 Cowie 认为搭配具有物理的、语法的和语义的 3 方面的特征[13],所以搭配现象是复杂的. Cowie 关注的是相关的句法和语义的搭配界定.

在研究汉语的搭配关系时还有其他的相关论述,孙新春(1992)在汉语词义研究的方法论中,总结了一种对汉语来说比较重要的语义搭配法,“语义搭配法就是通过一个词在与不同的词语形成不同的搭配关系来显示、判断词义的方法.”在他的文章中他还指出“语义搭配法对汉语来说具有特别重要的意义,因为汉语词没有形态,不像有形态的语言,一个词的具体意义和功能 总在它的外部形态上表现出来.”同时作者也指出了该方法的局限性.

2.2　搭配获取的方法和评价

2.2.1　手工搭配抽取

在计算机得到广泛应用之前,从大规模的语料库中自动抽取搭配只是理论上的. 所以之前搭配主要是通过编者的语感来发现获取,手工抽取搭配需要大量的人力和物力,并且能够处理的语料数量是有限的. 所以随着计算机的发展,自动抽取搭配成为研究和应用的必要选择.

2.2.2　自动搭配抽取

相对于手工搭配抽取,自动抽取搭配可以处理大规模的语料[14],以计算机程序为主,以人工校对为辅,可以很快地获得搭配候选集合,节省了人工认定的工作量.

自动搭配抽取一般的方法主要是将搭配的特征进行形式化、数量化,从而计算机可以自动判断搭配,得到最好的搭配结果. 自动抽取搭配主要分 3 步进行:(1)准备语料,根据研究的搭配抽取的目标选取合适的生语料或者经过基本处理过的标注语料. (2)根据语言学知识过滤候选的搭配集合. (3)抽取的搭配按照某个测度标准评价关联强度并排序. 一般搭配的测度可以选择一些统计指标,如对数似然度、卡平方和互信息作为两个词的关联强度评价[11].

根据 Dunning 在 1993 年提出的计算公式,对数似然度可以表示为:

$$\log\omega=\log L(\lambda_{12},\lambda_1,\rho)+\log L(\lambda_2-\lambda_{12},N-\lambda_1,\rho)-\log L(\lambda_{12},\lambda_1,\rho_1)-\log L(\lambda_2-\lambda_{12},N-\lambda_1,\rho_2),\quad(3)$$

其中,$L(k,n,x)=x^k(1-x)^{n-k}$,$\rho=\lambda_2/N$,$\rho_1=\lambda_{12}/\lambda_1$,$\rho_2=(\lambda_2-\lambda_{12})/(N-\lambda_1)$.

λ_1、λ_2、λ_{12}分别表示词语出现的次数和词语共现的次数. 对数似然比的值越大,两个词语成为搭配的可能性越大.

卡平方(χ^2)通过计算观测值和期望值之间差别的总和,测试词语 λ_1、λ_2 之间的关联性,公式可如下:

$$\chi^2 = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \tag{4}$$

式中,O_{ij}表示(i,j)的观测值,E_{ij}表示(i,j)的期望值. 卡平方值越大说明 λ_1、λ_2 之间的关联性越大.

互信息(Mutual Information,MI)[25]来自于信息论,用来测量两件事关联的信息量. 在搭配的研究中主要用来测试两个词语之间的关联强度,可以按照下列公式计算:

$$I(\lambda_1,\lambda_2) = \log 2 \frac{p(\lambda_1,\lambda_2)}{p(\lambda_1)\times p(\lambda_2)}, \tag{5}$$

$p(\lambda_1,\lambda_2)$表示 λ_1、λ_2 两个词在语句中共现的概率,$p(\lambda_1)$、$p(\lambda_2)$表示这两个词语在语句中各自出现的概率. 如果两个词语可以构成搭配 $p(\lambda_1,\lambda_2)>p(\lambda_1)\times p(\lambda_2)$,则 $I(\lambda_1,\lambda_2)>0$;如果两个词语没有紧密的联系,$p(\lambda_1,\lambda_2)\approx p(\lambda_1)\times p(\lambda_2)$,则 $I(\lambda_1,\lambda_2)\approx 0$;如果 $p(\lambda_1,\lambda_2)<p(\lambda_1)\times p(\lambda_2)$,$\lambda_1$、$\lambda_2$ 互补,则 $I(\lambda_1,\lambda_2)<0$.

搭配抽取方法的评价主要用根据手工答案作为标准搭配的集合,计算自动搭配抽取算法的准确率和召回率.

3 基于词类和搭配的文本聚类方法研究

3.1 方法概述

以往的研究主要是采用 N-gram 检索的方法来测文本中语句的相似度,即两个语句如果被测出出现某个相同的 N 元组,就认为两个句子中的这两个元组是相同的,赋值成 1,否则赋值为 0. 实际上,在自然语言中语句之间的相似度可以用 0 到 1 之间的数值来表示语句之间的关系亲近与否. 关系近的两个语句相似度应该接近于 1,关系远的应该更接近于 0. 本文利用语句中词语或者词类之间的搭配来探究语句之间的相似度,通过词类解决了词语的近义词或者同义词的现象,因为同义词被聚到同一个词类中,同一个词的不同意思被聚到不同的词类中.

搭配较之前的很多 N 元组的研究成果,它是更贴近于研究者关注的主题的词语搭配对,搭配还可以利用自身的特点去除更多冗余的词语从而加快程序的运行效率,提高结果的准确度. 第三节从语句中自动抽取的有效搭配,每条微博文本中都可以抽取出若干搭配对,对照两条微博之间搭配对的相似程度,从而计算两条微博的距离.

两条微博抽取的搭配例子如下:

Weibo1 (match1,match2)(match3,match4)(match1,match5)…

Weibo2 (match1,match2)(match6,match7)(match3,match5)…

其中 match1、match2、match3、match4、match5、match6、match7 可能是词条也可能是由词类模型构造的词类.

从两条微博中提取的有效搭配中可以看出(match1,match2)这一对搭配是相同的,但这两条微博之间的相似性如何的计算公式可参考 NIST(National Institute of Standard s and Technology)这个标准[15].

$$\mathrm{NIST} = \mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N} W_n \log R_n\right), \tag{6}$$

其中 BP 是 Brevity Penalty 的缩写,是指长度罚分. 假设 t 为被测文本的单词数,r 为参照文本的单词数. 如果 $t>r$,则 BP=1,否则 $\mathrm{BP}=e^{(1-r/t)}$. W_n 是各个搭配的权重系数. R_n 是搭配词组匹配时的分数,公式一般为:

$$R_n = \frac{\sum_{\mathrm{matches}\in(t\cap r)} \mathrm{Count}(\mathrm{matches})}{\sum_{\mathrm{matches}\in(t\cup r)} \mathrm{Count}(\mathrm{matches})}, \tag{7}$$

公式的分子是被测微博和参照微博共现的搭配词语的个数,分母是被测微博和参照微博所有非重复的搭配词语的个数.

本文采用向量空间模型(VSM)来表示文本. 在该模型中,每个微博文本被表示为向量空间中的一个向量,用 TF-IDF 来衡量每一个有效搭配对的权重.

本文采用了 K-Means 的算法进行文本聚类,K-Means 算法的基本思想如下:在给定的初始数据集中随机地选择 k 个数据对象作为 k 个组的初始中心点分别计算剩余的其他数据对象与各个组的初始中心点的距离,并根据距离进行归类,距离最近某中心点最近的数据对象则放到该中心点所在的组中,待所有数据对象全部归类后,再重新计算每个组的平均值作为新的中心点,重复迭代过程,直到完成指定的迭代次数,或聚类准则函数 E 收敛.

K-Means 算法的优点是聚类速度很快,但该算法也有缺点,如容易迭代过程中达到局部最优. 因此在进行开始聚类时如何确定数据的中心很重要,否则得不到想要的聚类效果.

3.2 方法详细设计

整个基于搭配的文本聚类算法主要包括以下几个步骤:

(1)微博文本预处理

本文使用的关于保险的微博语料是经过爬虫程序得到的有限条的微博语料[16],但由于程序的局限性,有时会有无关行业的语料被爬出,此时需要对爬出的程序进行人工校对,筛选出无关行业的语料或者重复的语料,从而提高语料的质量.

分词[21]是文本预处理中的重点,现在最主要的分词技术是字符串匹配分词方法和基于统计的分词方法,在理论的驱使下中科院计算所开发出了汉语词法分析器:ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System),该分析器可以进行汉语的分词、词性标注、新词识别和用户词典. 本文使用该开源系统来对微博语料进行分词.

对于一些跟保险行业的微博没有关系或者对理解保险行业不起作用的词语,如一些语气词还有一些频率过高或者过低的词语. 本文采用“停用词库”来处理停用词[22],停用词库中会去除一些超高频的词,如:“的/u”、“我/r”、“你/r”、“我们/r”等,这些词有很高的文档频率但反文档频率却很低[17]. 同时一些语气词,如:“恩/y”、“啊/y”、“呀/y”、“哇/y”、“啦/y”等也会被从语料中去除,从而减轻了后续算法的工作量.

(2)词类模型

词类模型可以解决文本中同义词或者多义词的现象,同义词被聚到同一个词类中,多义词将会根据词义归纳到不同的词类中,从而可以得到更加准确的文本相似度. 构造词类的主要步骤是词义归纳(Word Sense Induction,WSI)和词义聚类(Word Sense Clustering,WSC)[14].

①词义归纳(WSI)

WSI 是从大量的未标注语料中自动发现词义,经过国内外多名学者的研究,词义归纳算法有很多,但主要被分为 3 个类,基于特征向量的方法、基于图的方法和基于统计模型的方法. 本文以保险行业作为固定的主题,要获取与该主题相关的词义,先手工获取一部分词语的词义,用该词义个数作为先验,采用 Samuel Brody 提出的贝叶斯词义归纳(Bayesian Word Sense Induction)方法自动补充词义个数. 本文采用每条微博作为某个词语的上下文,直接采用 LDA(Latent Dirichlet Allocation)模型进行词义归纳. LDA 是典型的贝叶斯网络,该模型具有逻辑清楚的层次结构,用公式可表示,在文档中第 i 个词 w_i 的生成概率 $P(w_i)$

$$P(w_i)=\sum_{j=1}^{S}P(w_i|s_i=j)P(s_i=j). \tag{8}$$

$P(s_i=j)$ 表示文档包含某一词义的概率;$P(w_i|s_i=j)$ 表示文档中的词 w_i 出现在词义 s_i 下的概率. 根据 LDA 文本生成过程,也可以得到文档 d 中包含词语 w_i 的概率:

$$P(w_i|d)=\sum_{j=1}^{S}\varphi_{wi}^{j}\cdot\theta_j^d. \tag{9}$$

②词义聚类(WSC)

由于 LDA 模型只是针对词语的多义性的,没有考虑到同义性词语之间的关系,所以本文以微博中上下文词语作为特征词,$P(w_i|d)$ 作为特征词语的权重,对文本利用聚类算法进行聚类. 本文采用 K-Means 算法[14]进行聚类. 该聚类算法首先将样本词义作为一个类,经过不断迭代找到最大的类簇. 最后可以尽可能多的找到某一特征词的词义个数.

该模型最后用词类代替微博文本中的词语,即把文本表示在词类的空间上.

③搭配自动抽取算法

经过前面对微博文本的处理,此时需要对文本中的搭配进行自动抽取,主要步骤如下:

步骤 1:以微博中名词、动词、形容词和成语等作为节点词,微博文本中其他其他词语作为他们的搭配候选词. 文本中存在一些出现频率较高但对保险主题毫无意义的词语,应当作为停用词统计下来,并不参与以后的计算,从而不至于因为这些无意义的词来影响运算效率. 下表列出部分被停用的词.

表 1　部分停用词

Table 1　Part of stop words

词语义项	出现次数	词语义项	出现次数	词语义项	出现次数	词语义项	出现次数
的/u	16 421	个/q	1 001	要/v	1 549	在/p	2 050
了/y	2 264	就/d	1 650	什么	984	让/v	964
是/v	3 395	也/d	1 009	年/q	961	吗/y	919
要/v	1 549	和/c	995	来/v	1 053	人/n	3 031
在/p	2 500	你们/r	1 005	被/p	1 044	大/a	2 034
一个/m	1 004	都/d	1 511	给/p	1 218	买/v	1 986
我/r	4 513	还/d	1 531	去/v	1 061	#/n	1 005
你/r	1 347	能/v	1 042	到/v	2 306	@/n	2 877

步骤 2:根据程序得到每条微博中的排除停用词后的搭配,排除所有少于出现 3 次的低频候选搭配对. 如“[郝,汉]”和“[郝,求助近类]”

步骤 3:排除得到的搭配中节点词和候选词相同的搭配对儿,如:“[摩托车,摩托车]”.

步骤 4:设定对数似然比的一个阈值[18],根据微博文本中剩余的词语、词类的不同频次确认高于该指定阈值时被确认为有效搭配.

程序自动抽取搭配后的结果中包含大部分有效搭配并有一部分无效搭配没有被过滤. 列举部分有效搭配对和无效搭配对.

部分有效搭配:(人寿保险近类,骗人近类)、(细心近类,安泰保险近类)、(安泰近类,快近类)、(客户近类,喜欢近类)、(满意,阳光保险近类)、(理赔近类,好近类)、(投诉近类,人寿保险近类)、(人寿保险近类,犯贱)、(人寿保险近类,拐骗)、(平安保险近类,骗人近类)、(平安保险近类,喋喋不休)、(服务近类,诚恳)、(过分,喋喋不休)、(姗姗来迟,保险近类)、(保险近类,表扬近类)、(表扬近类,平安保险近类)、(保险近类,表扬近类)、(FUCK,华泰保险近类)、(平安保险近类,打扰近类).

部分无效搭配:(就业近类,很近类)、(城市,行囊)、(保障近类,阅读)、(国际,课)、(车胎,换近类)、(归来,技术近类)、(公共,变成)、(闪,现场近类)、(车辆近类,激增)、(公共,变成)、(承认近类,手)、(免,单证)、(以来,快近类)、(力,真心近类)、(首创,好近类)、(首创,喜欢近类)、(单证,理赔近类).

显然,经过本算法的步骤四,有些搭配被过滤掉,但无效搭配被抽取是在所难免的. 本文突出的行业是保险,对该行业的争议很大,从感情色彩上来说,有的用户信任保险,在微博中会用一些赞扬的词语,相反,有些会投诉保险,用一些贬义词表达憎恶,还有用户对保险不熟悉进行咨询等;从另一方面最让人关心的莫过于保险行业的服务和业务. 自动抽取的有效搭配被认定为对保险行业的赞扬、投诉等带有感情色彩的、带有咨询、疑问和质疑等疑问色彩的、突出保险行业服务或业务上的词语或者词类的搭配对. 无效搭配被认为是搭配中的两个词语或者词类,只有 1 个或者零个词语或词类跟保险行业有联系.

(3)文本聚类

使用向量空间模型(VSM)表示文本,计算每个抽取的有效搭配的 TF-IDF 的值,形成特征向量[21],该特征向量其实是 1 个二维矩阵,即为搭配在文本聚类算法中的权值 $W[j][i]$. 统计每个搭配在每一个微博文本中的频率 $\mathrm{TF}(i,j)$ 和反文档频率 $\mathrm{IDF}(i)$. 则权值的公式为:

$$W[j][i]=\mathrm{TF}(i,j)*\mathrm{IDF}(i),$$

其中 i 是每条微博文本中第 i 个有效搭配对,j 表示第 j 条微博文本.

微博文本间的相似度采用 NIST(National Institute of Standards and Technology)[22]标准计算.

本文使用的 K-Means 算法流程如下:

①随机选取 k 条微博文本生成 k 个聚类,这 k 条文本分别分别对应 k 个聚类的聚类中心点为 $\{\mu_1,\mu_2,$

$\mu_3,\cdots,\mu_k\}$. 第 j 条微博文本的有效搭配可表示 MircoText$[J]=\{(m_m,m_n)\mid m,n\geqslant 1\}$.

②利用 NIST 标准计算每条微博文本与每个聚类中心对应微博的相似度,根据与聚类中心的相似程度将其分配到最近的 1 类中,得到新的聚类结果 $\{U_1,U_2,U_3,\cdots,U_k\}$,其中 $U_i(1\leqslant i\leqslant k)$ 是若干微博文本的集合.

③重新计算每个簇 $U_i(1\leqslant i\leqslant k)$ 的中心 $\mu_i'=\dfrac{\sum_{j=1}^{m}c_j}{\sum_{j=1}^{m}j}$,其中 m 是集合 $U_i(1\leqslant i\leqslant k)$ 中数据的总数,c_j 是指 $U_i(1\leqslant i\leqslant k)$ 中第 j 个数据.

④当下面的准则函数收敛时,停止迭代,否则重复第②步和第③步.

判断是否停止迭代的准则函数可表示为:$E(U,\mu)=\sum_{i=1}^{m}\|U_i-\mu_i\|^2$.

E 函数[23]表示每条微博文本所属类别到其质心的距离平方和. 假设当前 E 没有达到最小值,那么首先可以固定每个类的质心 μ_i,调整每个样例的所属的类别 U_i 来让 E 函数减少,同样,固定 U_i,调整每个类的质心 μ_i 也可以使 E 减小. 这两个过程就是内循环中使 E 单调递减的过程.

4 实验结果及分析

为了验证基于搭配的微博文本聚类的有效性,本文收集了新浪微博上的关于保险行业的 5 000 条数据,经过文本预处理后[24],分别用传统、基于词类和基于词类和搭配的 3 种文本聚类方法进行了实验,得到 3 种聚类结果. 结果如表 2 所示.

表 2 3 种聚类方法结果

Table 2 The results of three clustering methods

聚类方法	聚类后类别总数	多个元素一个簇	一个元素一个簇
传统文本聚类方法	1 124	603	521
利用词类文本聚类方法	1 053	612	441
基于词类和搭配文本聚类方法	876	623	253

从 3 种聚类方法的结果可以得到:在考查保险行业的微博文本 5 000 条时,文本利用词类进行文本聚类总体性能比传统文本聚类方法提高了 6.3%,基于词类和搭配的文本聚类总簇数比利用词类聚类的总体性能提升 16.8%. 在相同主题下,聚类后类别总数减少在一定程度上说明了本文方法更适合应用在微博文本,接着我们用计算聚类效果的评价指标进一步验证基于搭配的微博文本聚类方法的有效性.

聚类效果的评价指标[18]通常用准确率、召回率和 F-Measure. 人工标注的数据主题 $p=\{p_1,p_2,p_3,\cdots,p_j,\cdots,p_n\}$,聚类的簇 $c=\{c_1,c_2,c_3,\cdots,c_i,\cdots,c_m\}$. 准确率、召回率和 F-Measure 分别可表示为:

$$\text{Precision}(p,c)=\frac{|p\cap c|}{|c|},\tag{10}$$

$$\text{Recall}(p,c)=\frac{|p\cap c|}{|p|},\tag{11}$$

$$F(p,c)=\frac{2P(p,c)*R(p,c)}{P(p,c)+R(p,c)},\tag{12}$$

经过多名人员的手工标注,微博数据可以聚类为 937 个类簇,其中一个元素一个簇的是 285 个. 可以分别得到聚类方法一和聚类方法二的准确率、召回率和 F 值.

表 3 两种聚类方法结果评价指标

Table 3 The result's evaluation indicator of two clustering methods

聚类方法	总簇数	与人工相同簇	P 值(准确率)/%	R 值(召回率)/%	F-Measure 值/%
利用词类文本聚类方法	1 053	489	46.43	52.19	49.14
基于词类和搭配文本聚类方法	876	667	76.14	71.18	75.58

从实验结果可以看出,本文提出的基于搭配的文本聚类比直接聚类的 F-Measure 值高,本文的聚类方法效果更好.

该方法可以处理短文本的稀疏数据[25],但从实验结果看出,对于 5 000 条或者数据量更大的微博短文本准确率、召回率和 F 值相比较而言是低的. 分析指标低的主要原因有两个:

(1)词义归纳(WSI)不完善:由于汉语的表达方式多种多样,微博中用户可以用任何的表达方式来表达自己的感情[3]. 用 LDA 模型在词义归纳的过程中,某些词语的词义没有被归纳,从而词类中没有囊括该项词义,构成了遗漏.

(2)K-Means 聚类方法局限性:K-Means 算法有很明显的优点,即聚类速度很快,但它也有缺点,在迭代过程中容易达到局部最优,因此在选择数据中心时要取尽量小的偏差,否则聚类效果将前功尽弃,从而会影响聚类的最终结果. 运用到微博短文本上,聚类算法需要进一步的改进.

5 结束语

本文针对微博短文本提出了基于词类和搭配的文本聚类方法,该方法主要把词类模型和词语搭配自动抽取运用到文本聚类中,最后的文本聚类算法是利用 K-Means 算法. 词类模型主要是根据区分词语的词义把相近的词语聚为一类,同一个词的不同词义类别聚到不同的类中,所以解决了文本中的同义词和多义词的问题,但由于微博文本数据量过大,有一些和保险主题不相关的词语并没有被运用到词类模型中,运用这种模型不仅减少冗余词语的破坏更突出了保险这个主题,然后用程序实现自动抽取文本[26]中的有效搭配对儿,最后利用文本的相似度对文本进行聚类. 实验结果表明该方法准确率、召回率和 F 值都较高,证明了其有效性.

该方法是运用到大量微博短文本中的,虽然相较以前的研究成果指标有所提高,但仍没有达到理想的结果. 今后的研究重点是逐步改进 K-Means 聚类算法,使之更适用于大量短文本数据中. 另外进一步改进词类模型自动获取词语的词义的个数进行词义归纳.

[参考文献]

[1] 李勇,张克亮,李伟刚. 基于微博的网络舆情分析系统设计[J]. 计算技术与自动化,2013,32(2):2-5.

[2] 张洋,何楚杰,段俊文. 微博舆情热点分析系统设计研究[J]. 信息网络安全,2012(9):60-63.

[3] 唐国瑜,夏云庆,张民. 基于词义类簇的文本聚类[J]. 中文信息学报,2013,27(3):114-118.

[4] 董婧灵. 基于 LDA 模型的文本聚类研究[D]. 武汉:华中师范大学计算机科学系,2012.

[5] 石晶,李万龙. 基于 LDA 模型的主题词抽取方法[J]. 计算机工程,2010,39(19):81-83.

[6] 陈慧,石冰. 基于贝叶斯模型的微博虚假话题数据分析研究[D]. 山东:山东大学计算机科学与技术学院,2013.

[7] Pessiot J,Kim Y,Amini M,et al. Improving document clustering in a learned concet space[J]. Information Processing and Management,2010,46:180-192.

[8] Dhillon S. Co-clustering document and words using bipartite spectral graph partitioning[C]//UT CS Technical Report. Austin,2001:269-274.

[9] 朱鑫,词语搭配自动抽取方法对比研究[D]. 大连:大连海事大学计算机科学与技术学院,2010.

[10] 孙茂松,黄昌宁,方捷. 汉语搭配定量分析初探[J]. 中国语文,1997(1):29-38.

[11] 邓耀臣,王同顺. 词语搭配抽取的统计方法及计算机实现[J]. 外语电化教学,2005,105:25-26.

[12] 郎需超. 基于 R 值的汉语搭配抽取[D]. 北京:北京邮电大学计算机科学与技术学院,2012.

[13] Cowie A P, Mackin R, McCaig I R. Oxford Dictionary of Current Idiomatic English [M]. London: Oxford University Press,1975.

[14] Brody S,M Lapata. Bayesian word sense induction[C]//Proc of EACL. Bergen,Norway:European Chapter of the Association for Computational Linguistics,2009:101-113.

[15] 王金铨,梁茂成,俞洪亮. 基于 N-gram 和向量空间模型的语句相似度研究[J]. 现代外语,2007,30(4):406-412.

[16] 曾星宇,李淑琴,陈斌. 基于微博文本的舆情分析和研究[J]. 信息技术与信息化,2014(1):86-87.

[17] 林达真,面向博客的舆情分析若干关键技术研究[D]. 厦门:厦门大学计算机科学系,2012.

[18] 曲维光,陈小荷,吉根林. 基于框架的词语搭配自动抽取方法[J]. 计算机工程,2004,30(23):22-24.

[19] Tang G,Xia Y,Zhang M,et al. 2011 CLGVSM:adapting generalized vector space model to cross-lingual document clustering[C]// Proc of IJCNLP,Hainan Island:Springer,2010:578-588.
[20] Steinbach M,Karypis G,Kumar V. A comparison of document clustering techniques[C]//KDD Workshop on Text Mining. Boston,2000:368-503.
[21] 楼佳. 中文文本聚类的评价与改进研究[D]. 杭州:杭州电子科技大学计算机学院,2009.
[22] 刘远超,王晓龙,徐志明. 文档聚类综述[J]. 中文信息学报,2005,20(3):57-61.
[23] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京:中国科学院计算技术研究所,2005.
[24] 李勇,张克亮,李伟刚. 基于微博的网络舆情分析系统设计[J]. 计算机技术与自动化,2013,32(2):123-127.
[25] 时睿,面向短文本的网络舆情分析[D]. 西安:西安电子科技大学电子工程学院,2012.
[26] 陈雅菊,现代汉语词语搭配的自动抽取方法[D]. 上海:华东师范大学软件学院,2005.

[责任编辑:顾晓天]

(上接第 56 页)

[7] Thangavel K,Pethalakshmi A. Dimensionality reduction based on rough set theory:a review[J]. Applied Soft Computing,2009,9(1):1-12.
[8] 林俊伟,叶东毅. 基于领域辨识矩阵的属性约简增量式算法[J]. 计算机应用,2009,29(11):119-121
[9] Hu F,Wang G Y,Huang H,et al. Incremental attribute reduction based on element arysets[C]//Proceedings of the 10th International Conference on Rough Sets,Fuzzy Sets,Data Mining,and Granular Computing. Regina,2005:183-193
[10] 梁吉业,魏巍,钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践,2008,28(4):81-89
[11] Guoyin W,Yiyu Y,Hong Y. A survey on rough set theory and applications[J]. Chinese Journal of Computers,2009,32(7):1 229-1 246.
[12] Yu H,Liu Z,Wang G. An automatic method to determine the number of clusters using decision-theoretic rough set[J]. International Journal of Approximate Reasoning,2014,55(1):101-115.
[13] Jia X,Liao W,Tang Z,et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences,2013,219:151-167.
[14] Chen H,Li T,Ruan D,et al. A rough-set-based incremental approach for updating approximations under dynamic maintenance environments[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(2):274-284.

[责任编辑:顾晓天]