(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110941713 B
(45)授权公告日 2023.12.22

(21)申请号 201811107536.8

(22)申请日 2018.09.21

(65)同一申请的已公布的文献号
申请公布号 CN 110941713 A

(43)申请公布日 2020.03.31

(73)专利权人 上海仪电(集团)有限公司中央研究院
地址 200233 上海市徐汇区虹漕路39号4号楼6层

(72)发明人 张鹏飞

(74)专利代理机构 上海科盛知识产权代理有限公司 31225
专利代理师 翁惠瑜

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/332(2019.01)

(56)对比文件
CN 102902700 A,2013.01.30
CN 107239529 A,2017.10.10
CN 107169001 A,2017.09.15
CN 105930360 A,2016.09.07
CN 102023967 A,2011.04.20
CN 105718444 A,2016.06.29
CN 105975478 A,2016.09.28
杨春明;何天翔.元搜索引擎的结果去重及排序研究.软件.2012,(06),正文第1节.
姚立.基于主题模型的改进随机森林算法在文本分类中的应用.计算机应用与软件.2017,(第08期),正文第1-3节.
姚立.基于主题模型的改进随机森林算法在文本分类中的应用.计算机应用与软件.2017,(第08期),正文第1-3节.

审查员 王红微

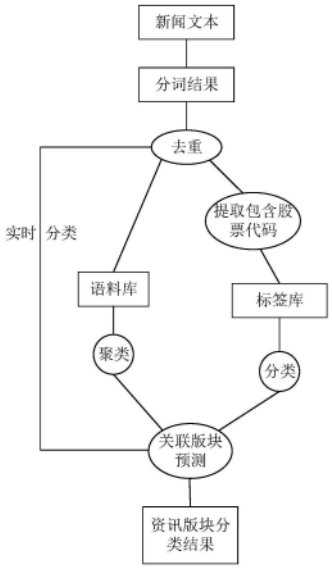
权利要求书1页 说明书4页 附图1页

(54)发明名称

基于主题模型的自优化金融资讯版块分类方法

(57)摘要

本发明涉及一种基于主题模型的自优化金融资讯版块分类方法,包括以下步骤:1)获取网络抓取的股票证券行业资讯文本,对所述文本进行分词处理,获取与所述文本对应的词汇;2)判断是否需要进行去重处理,若是,则去重后返回步骤1),若否,则执行步骤3);3)从所述词汇中提取股票名称和/或股票代码,记录每一股票名称或股票代码对应的股票版块,形成与所述文本对应的版块标签;4)基于所述词汇利用自动更新的关联版块预测模型获得关联预测概率;5)基于版块标签和关联预测概率获得所述文本在某个版块分类上的评分,以评分最高的版块分类作为推荐分类。与现有技术相比,本发明具有良好的自动扩展能力与随时间自动优化的能力。



CN 110941713 B

1.一种基于主题模型的自优化金融资讯版块分类方法,其特征在于,包括以下步骤:

1)获取网络抓取的股票证券行业资讯文本,对所述文本进行分词处理,获取与所述文本对应的词汇;

2)判断是否需要进行去重处理,若是,则去重后返回步骤1),若否,则执行步骤3);

3)从所述词汇中提取股票名称和/或股票代码,记录每一股票名称或股票代码对应的股票版块,形成与所述文本对应的版块标签;

4)基于所述词汇利用自动更新的关联版块预测模型获得关联预测概率;

5)基于步骤4)的版块标签和步骤5)的关联预测概率获得所述文本在某个版块分类上的评分,以评分最高的版块分类作为推荐分类;

所述关联版块预测模型的训练优化具体为:

101)以历史文本及其词汇作为语料库,对语料库中的每个词汇w随机赋予一个topic编号;

102)扫描语料库,对每个词汇w,使用Gibbs Sampling公式对其采样,更新其topic编号,直至Gibbs Sampling收敛;

103)建立语料库的topic-word共现频率矩阵;

104)以所述topic-word共现频率矩阵作为训练数据,以版块标签作为训练目标,进行关联版块预测模型的分类训练优化;

所述分类训练优化基于随机森林实现,所述随机森林中的参数k通过以下公式选择:

$k=\log_2 d+\log_2 c+1$

式中,d为标签库中资讯总数,c为标签库中的分类数量;

所述文本在某个版块分类上的评分的计算公式为:

$$P_i=\frac{k_i}{k}\times 1.2+RF_i$$

式中,k_i为所述文本在第i个版块分类所属股票名称或股票代码出现的次数,k为所有股票名称或股票代码在该文本中出现的次数,RF_i为所述文本在第i个版块分类上的关联预测概率。

2.根据权利要求1所述的基于主题模型的自优化金融资讯版块分类方法,其特征在于,所述去重处理具体为:

采用TF-IDF向量计算当前文本与历史文本的相似度,删除相似度大于设定阈值的文本。

3.根据权利要求2所述的基于主题模型的自优化金融资讯版块分类方法,其特征在于,所述历史文本为从当前文本接收时间起过去1小时内的文本。

4.根据权利要求1所述的基于主题模型的自优化金融资讯版块分类方法,其特征在于,所述关联版块预测模型按设定周期进行训练优化。

基于主题模型的自优化金融资讯版块分类方法

技术领域

[0001] 本发明涉及金融数据处理技术领域，尤其是涉及一种基于主题模型的自优化金融资讯版块分类方法。

背景技术

[0002] 在金融证券行业，资讯消息对于从业人员是不可忽视的重要参考信息来源，因此消息的及时性、准确性、可靠性成为了行业从业人员非常关心的问题。随着信息时代的到来，资讯的获取途径也逐渐向网络化、信息化进行转移，越来越多的信息化手段能够辅助行业人员进行咨询的获取、汇聚。但相应的，在信息的爆炸时代，如何有效的筛选、甄别和分类获取的咨询，对于行业人员第一时间聚焦到有效、相关的咨询信息具有重要意义。在这其中，咨询文本对应的股票版块分类是一个最为迫切、常见的需求。

[0003] 针对网络各类信息源获取/爬取的咨询文本信息，目前也有几种基于自然语言处理技术的文本分类方法可以辅助人工进行文本分类，但是目前绝大多数分类算法都需要来源于数量庞大的具有标注的数据，而且随着技术、市场的不断变化，文本的分类规则和具体标记有可能也会发生相应的改变，因此很多基于历史上的人工标记的咨询版块分类数据训练得出的分类算法，并不能很好的适用于新的咨询。

发明内容

[0004] 本发明的目的就是为了克服上述现有技术存在的缺陷而提供一种基于主题模型的自优化金融资讯版块分类方法。

[0005] 本发明的目的可以通过以下技术方案来实现：

[0006] 一种基于主题模型的自优化金融资讯版块分类方法，包括以下步骤：

[0007] 1)获取网络抓取的股票证券行业资讯文本，对所述文本进行分词处理，获取与所述文本对应的词汇；

[0008] 2)判断是否需要进行去重处理，若是，则去重后返回步骤1)，若否，则执行步骤3)；

[0009] 3)从所述词汇中提取股票名称和/或股票代码，记录每一股票名称或股票代码对应的股票版块，形成与所述文本对应的版块标签；

[0010] 4)基于所述词汇利用自动更新的关联版块预测模型获得关联预测概率；

[0011] 5)基于步骤4)的版块标签和步骤5)的关联预测概率获得所述文本在某个版块分类上的评分，以评分最高的版块分类作为推荐分类。

[0012] 进一步地，所述去重处理具体为：

[0013] 采用TF-IDF向量计算当前文本与历史文本的相似度，删除相似度大于设定阈值的文本。

[0014] 进一步地，所述历史文本为从当前文本接收时间起过去1小时内的文本。

[0015] 进一步地，所述关联版块预测模型的训练优化具体为：

[0016] 101)以历史文本及其词汇作为语料库，对语料库中的每个词汇w随机赋予一个

topic编号；

[0017] 102)扫描语料库，对每个词汇w，使用Gibbs Sampling公式对其采样，更新其topic编号，直至Gibbs Sampling收敛；

[0018] 103)建立语料库的topic-word共现频率矩阵；

[0019] 104)以所述topic-word共现频率矩阵作为训练数据，以版块标签作为训练目标，进行关联版块预测模型的分类训练优化。

[0020] 进一步地，所述分类训练优化基于随机森林实现，所述随机森林中的参数k通过以下公式选择：

[0021] $k=\log_2 d+\log_2 c+1$

[0022] 式中，d为标签库中资讯总数，c为标签库中的分类数量。

[0023] 进一步地，所述关联版块预测模型按设定周期进行训练优化。

[0024] 进一步地，所述文本在某个版块分类上的评分的计算公式为：

[0025] $$P_i=\frac{k_i}{k}\times 1.2+RF_i$$

[0026] 式中，k_i为所述文本在第i个版块分类所属股票名称或股票代码出现的次数，k为所有股票名称或股票代码在该文本中出现的次数，RF_i为所述文本在第i个版块分类上的关联预测概率。

[0027] 与现有技术相比，本发明通过持续抓取互联网上的金融资讯，自动形成不断累积和演化的语料库与标签库，通过语料库与标签库定期的训练生成符合时代变化和技术更新的资讯股票版块关联预测模型，解决金融类资讯版块分类问题中，标记数据量小、难以生成，且难以随时代技术的发展而变化的问题。相对于传统的标记数据训练方法，该方法具有良好的自动扩展能力与随时间自动优化的能力。

附图说明

[0028] 图1为本发明的流程示意图。

具体实施方式

[0029] 下面结合附图和具体实施例对本发明进行详细说明。本实施例以本发明技术方案为前提进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

[0030] 本发明提供一种基于主题模型的自优化金融资讯版块分类方法，包括以下步骤：1)获取网络抓取的股票证券行业资讯文本，对所述文本进行分词处理，获取与所述文本对应的词汇；2)判断是否需要进行去重处理，若是，则去重后返回步骤1)，若否，则执行步骤3)；3)从所述词汇中提取股票名称和/或股票代码，记录每一股票名称或股票代码对应的股票版块，形成与所述文本对应的版块标签；4)基于所述词汇利用自动更新的关联版块预测模型获得关联预测概率；5)基于步骤4)的版块标签和步骤5)的关联预测概率获得所述文本在某个版块分类上的评分，以评分最高的版块分类作为推荐分类。

[0031] 如图1所示，本系统处理流程如下：

[0032] 1)针对网络抓取的股票证券行业资讯文本，首先利用成熟的中英文分词技术，结

合金融行业特定词典,进行分词处理。

[0033] 2)分词后的文本数据与历史数据库对比,排除重复多余的资讯,对于去重方法,本发明采用TF-IDF向量计算两条文本的相似程度,对于相似度过大的两条咨询,认为是重复咨询,计算方式如下:

[0034] $$TFIDF_{ij}=\frac{n_{ij}}{\sum_k n_{kj}}\times \log j\frac{|D|}{|j:t_i\in d_j|}$$

[0035] 上式为单词i在资讯j中的TFIDF值。其中TF计算单词出现次数与资讯分词后单词总数之比,IDF为全库单词数与包含单词i的资讯数量比值的对数。

[0036] 值得注意的是,由于资讯库在持续累积,因此本方法仅对过去1小时内抓取的资讯进行IDF库计算,而每条新抓取的资讯,均与1小时内所有其他资讯进行逐一比对TFIDF向量的相似度,最终重复判定如下:

[0037] $$similarity=\cos(\theta)=\frac{A\bullet B}{\|A\|\|B\|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}}$$

[0038] 当上述相似度大于给定重复判定阈值时,讲判定为该两条资讯重复,会把时间更新的一条标记为重复新闻。

[0039] 3)去重之后的有效文本存储入语料库,作为文本主题聚类训练的储备数据。语料库储存所有爬取的资讯的文本,每条新闻的所有文字为1条语料。

[0040] 4)对于资讯中可能存在的股票名称、股票代码进行提取,提取出的对应股票检查其所属的股票版块(每只股票所属版块通过证券交易所数据获取,并经过从业人员验证认可,该映射关系变动较小,无需频繁更新)。

[0041] 5)对于4)中的结果,当一条资讯包含的所有股票均属于同一版块的情况下(排除一些无主题股评情况的影响),将该资讯标记为其对应的版块标签,存入标签库。标签库包含了新闻的文本以及其对应的板块分类标签(如“汽车行业”等),标签库每条数据格式为:<新闻文本,板块分类标签>。

[0042] 6)同时,标签库也将包含人工处理和标定的资讯分类结果。

[0043] 7)对于步骤3)中存储的文本语料库数据,进行LDA主题聚类运算,具体算法如下:

[0044] • 1.对语料库中的每篇文档中的每个词汇w,随机的赋予一个topic编号z;

[0045] • 2.重新扫描语料库,对每个词w,使用Gibbs Sampling公式对其采样,求出它的topic,在语料中更新;

[0046] • 3.重复步骤2,直到Gibbs Sampling收敛;

[0047] • 4.统计语料库的topic-word共现频率矩阵,该矩阵就是LDA的模型。

[0048] 值得注意的是,上述LDA模型生成过程可能持续时间较长,且过程中可能加入新的爬取资讯,为保证模型收敛,我们选取系统数据进入较少的时段(交易时间以外的时段新闻资讯产生速率会降低),并且在LDA训练开始后对于新加入的资讯不予计入。上述LDA训练过程每天进行一次即可,目的是能够使模型随着资讯的积累不断的优化。

[0049] 8)对于标签库中的数据,进行LDA模型分析后,形成单条资讯属于各不同主题的概率分布,步骤如下所示:

[0050] 1.对当前文档中的每个单词专栏w随机初始化一个topic编号z；

[0051] 2.使用Gibbs Sampling公式，对每个w重新采样其topic；

[0052] 3.重复以上过程，直至Gibbs Sampling收敛；

[0053] 4.统计文档中的topic分布。

[0054] 进一步的，将得到的每条资讯对应的topic分布形成的数据向量作为训练数据，利用资讯的版块标记作为训练目标，进行随机森林(或其他分类算法)的分类训练和优化。在随机森林中，对于决策树的每个结点，显示从当前节点的全部属性集合中随机选择一个包含k个属性的子集，之后再从这个子集中选择一个最优的划分属性。而在金融资讯中，训练集是随着时间增长的，为了平衡训练集的大小、样本相关度变化以及资讯分类数量的变化，本发明创新性的定义了k的选取方式如下：

[0055] $k=\log_2 d+\log_2 c+1$

[0056] 其中，d为标签库中资讯总数，c为标签库中的分类数量。可见，随着数据总量和分类数量的增长，k会有相应的变化，但是并不会增长很快。

[0057] 9)有了8)中的预测模型，对于新的一条资讯，经分词和去重处理后，就可以利用其LDA特征(即topic概率分布)进行分类。再结合其包含股票所属版块与步骤8)中的模型预测版块结果进行该资讯的关联版块推荐，这里给出资讯在某个分类上的关联度评分算法，根据该算法计算得到的关联度最高的分类，作为推荐分类给出。

[0058] $$P_i=\frac{k_i}{k}\times 1.2+RF_i$$

[0059] 上式中，ki为该资讯第i个分类所属股票代码出现的次数，k为所有股票代码在该资讯中出现次数，RF为对应分类的关联预测概率。

[0060] 以上详细描述了本发明的较佳具体实施例。应当理解，本领域的普通技术人员无需创造性劳动就可以根据本发明的构思作出诸多修改和变化。因此，凡本技术领域中技术人员依本发明的构思在现有技术的基础上通过逻辑分析、推理或者有限的实验可以得到的技术方案，皆应在由权利要求书所确定的保护范围内。

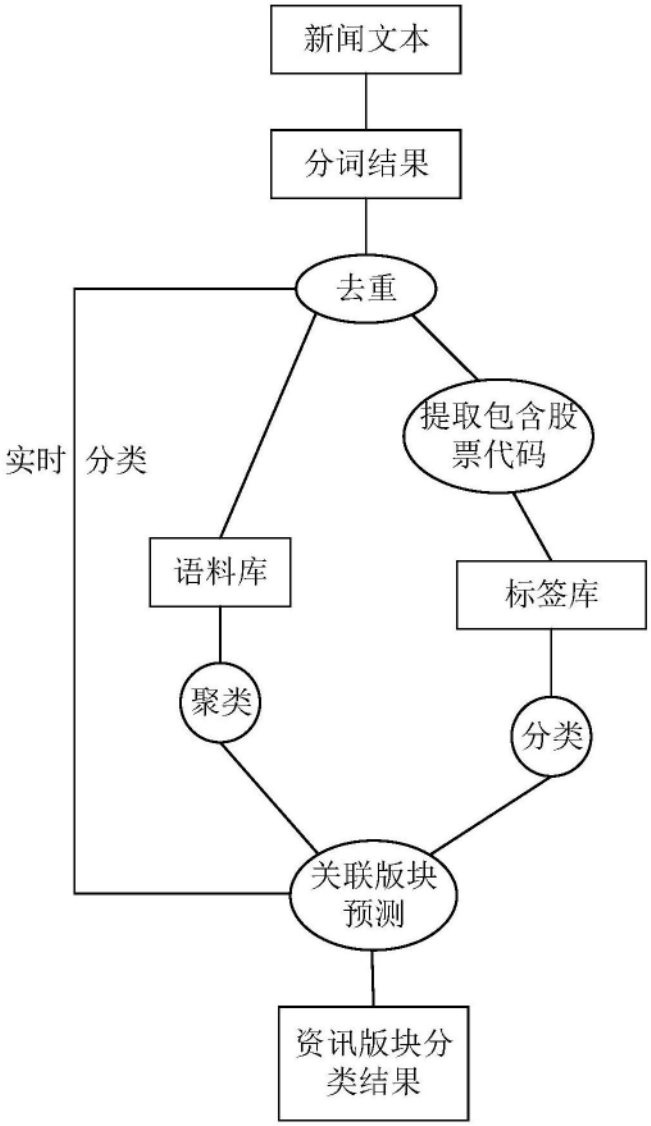


图1