(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103154991 B
(45)授权公告日 2017.03.29

(21)申请号 201180040396.7

(22)申请日 2011.07.21

(65)同一申请的已公布的文献号
申请公布号 CN 103154991 A

(43)申请公布日 2013.06.12

(30)优先权数据
12/842,440 2010.07.23 US

(85)PCT国际申请进入国家阶段日
2013.02.20

(86)PCT国际申请的申请数据
PCT/US2011/044830 2011.07.21

(87)PCT国际申请的公布数据
WO2012/012623 EN 2012.01.26

(73)专利权人 汤森路透环球资源公司
地址 瑞士巴尔

(72)发明人 莱恩·D·罗塞 乔治·P·邦尼

(74)专利代理机构 北京市浩天知识产权代理事务所(普通合伙) 11276
代理人 宋菲 刘云贵

(51)Int.Cl.
G06Q 40/00(2012.01)

(56)对比文件
US 2005/0071217 A1,2005.03.31,说明书第5、7、9、30、35、51段.

审查员 刘剑

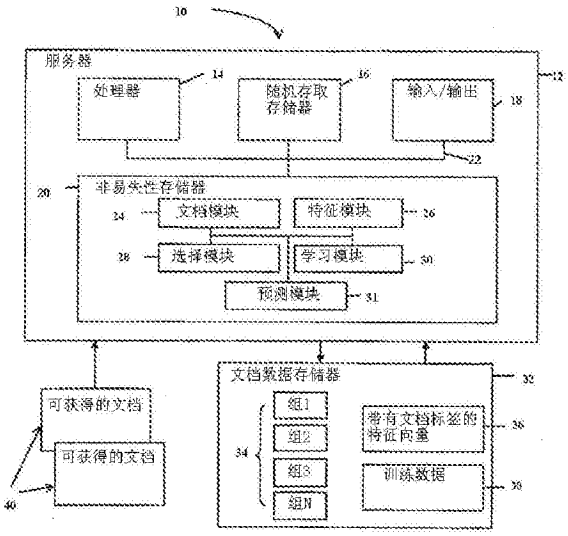
权利要求书3页 说明书8页 附图3页

(54)发明名称

信用风险采集

(57)摘要

本发明公开了使用各种数据源来开发和实施信用风险模型的系统和技术,这些数据源包括价格数据、财会比、ESG(环境、社会和政府)数据和文字数据。每种数据源提供与工厂或公司等实体的健康状况有关的唯一并且独特的信息。这些系统和技术组合不同来源的信息,产生特别强大的信号。这些系统和技术可以用来预测许多事件,这些事件包括但不限于拖欠债务或破产的概率、违约损失率、评级机构评级变化的概率,和股本价格变动的概率。



CN 103154991 B

1.一种判断一家公司的信用风险的方法，包括：

为第一组文档中包括的每个文档分配对象描述符，该对象描述符依据于历史事件或量化量度，其中，通过使日期值、实体识别符和标签值与该第一组文档中的每个文档相关联来产生该对象描述符，该日期值表示首次获得每个文档的时间段；

为该第一组文档中的每个文档分配至少一个特征向量，该至少一个特征向量是数值特征的N维向量，这些数值特征各自表示识别出来的文字内容、识别出来的语义内容，或这两种内容的组合；

判断多个所分配的对象描述符与多个所分配的特征向量之间的关系；

依据这种关系为第二组文档中不包括在该第一组文档中的每个文档分配预测描述符，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件；以及

依据该第二组文档中的至少一个预测描述符产生信号；

其中，所述方法还包括依据与该第一组文档中的文档相关联的文字内容、元数据或指示符产生该至少一个特征向量；

从该第一组文档中随机选择预先确定的数目的文档，形成多个文档子集，该多个子集中的每个文档包括该至少一个特征向量和相关联的第一标签值，或者不同于该第一标签值的相关联的第二标签值；

计算该至少一个特征向量的每个特征的量度值，方法是通过判断该特征是描述该相关联的第一标签值还是该相关联的第二标签值；以及

计算每个子集的每个特征的秩值，方法是通过汇集该多个子集中的每个特征的多个计算得来的量度值。

2.根据权利要求1所述的方法，进一步包括汇集多个文档，形成该第一组文档中的至少一个文档。

3.根据权利要求2所述的方法，其中汇集该多个文档包括将与该多个文档相关联的多个日期值与预先确定的时间段进行比较。

4.根据权利要求1所述的方法，包括为该第一组文档中的文档群组分配该至少一个特征向量。

5.根据权利要求1所述的方法，其中该量化量度是感情评级、语气评级、公司事件指示符、语言分析指示符、统计分析指示符或者以上内容的组合中的一项。

6.根据权利要求1所述的方法，该标签值表示该历史事件或该量化量度。

7.根据权利要求6所述的方法，其中关联该日期值包括判断首次获得该文档的时间段。

8.根据权利要求6所述的方法，包括使多个标签值与该第一组文档中的至少一个文档相关联。

9.根据权利要求6所述的方法，其中该标签值表示多个历史事件或多个量化量度。

10.根据权利要求6所述的方法，包括：

判断在该第一组文档的文档中是否识别出多个实体；以及

依据判断结果为该文档分配多个对象识别符，

其中该多个对象识别符中的每一个对象识别符对应于该多个实体中的一个实体，并且该文档的标签依据以下两项来修改：1)对应于该一个实体的该历史事件或该量化量度；和

2)对应于该一个实体的该文档的相关度分数。

11.根据权利要求1所述的方法,其中产生该至少一个特征向量包括将该文档剖析成一组词语或短语。

12.根据权利要求11所述的方法,包括在产生该至少一个特征向量之前,先从该组词语或短语中去掉预先识别的词语。

13.根据权利要求11所述的方法,包括在产生该至少一个特征向量之前,先使用提取词干算法提取该组词语或短语中包括的一个或一个以上词语的词干。

14.根据权利要求13所述的方法,其中该提取词干算法是波特提取词干算法。

15.根据权利要求1所述的方法,进一步包括:

将该计算得来的秩值与预先确定的阈值进行比较;以及

依据比较结果,将包括具有相关联的第一标签值或第二标签值的多个有阶特征的特征向量提供给机器学习模块,用于判断该关系。

16.根据权利要求15所述的方法,其中该机器学习模块使用回归算法、支持向量机(SVM)、神经网络或决策树算法中的一项。

17.根据权利要求15所述的方法,进一步包括交叉验证该多个子集的该判断的关系,从而判断最佳数目的特征向量,用于提供给该机器学习模块。

18.根据权利要求17所述的方法,进一步包括使用该机器学习模块产生该第二组文档中的每个文档的该预测描述符。

19.根据权利要求18所述的方法,其中该预测描述符依据于最新接收到的文档或在预先确定的时间间隔中接收到的文档的汇总。

20.一种判断一家公司的信用风险的方法,包括:

为第一文档分配预测描述符,该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件;以及

依据该预测描述符产生实体信用风险信号;

其中,该预测描述符依据于对第一训练文档与第二训练文档之间的关系的判断;

为该第一训练文档和该第二训练文档分配对象描述符,该对象描述符依据于历史事件或量化量度,其中,通过使日期值、实体识别符和标签值分别与该第一训练文档和该第二训练文档相关联来产生对象描述符,日期值表示首次获得文档的时间段;以及

为该第一训练文档和该第二训练文档分配至少一个特征向量,该至少一个特征向量是数值特征的N维向量,这些数值特征各自表示识别出来的文字内容、识别出来的语义内容,或这两种内容的组合。

21.根据权利要求20所述的方法,进一步包括判断多个对象描述符与多个特征向量之间的关系。

22.一种判断一家公司的信用风险的装置,包括:

用于为第一文档分配预测描述符的装置,该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件;

用于依据该预测描述符产生实体信用风险信号的装置;

其中该预测描述符依据:第一训练文档,该第一训练文档具有对应的第一对象描述符

和对应的第一存储的数值表示，该对应的第一对象描述符和该对应的第一存储的数值表示具有第一预定义关系，其中，通过使日期值、实体识别符和标签值与该第一训练文档相关联来产生该第一对象描述符，该日期值表示首次获得第一训练文档的时间段。

23.根据权利要求22所述的装置，其中该预测描述符依据于第二训练文档，该第二训练文档具有对应的第二对象描述符和对应的第二存储的数值表示，该对应的第二对象描述符和该对应的第二存储的数值表示具有第二预定义关系。

24.一种判断一家公司的信用风险的系统，包括：

数据存储器，包括第一组文档和第二组文档；

服务器，包括处理器和存储器，该存储器存储响应于接收到访问服务的请求；

所述处理器用于：

为该第二组文档中的第一文档分配预测描述符，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件，该预测描述符依据于该第一组文档中的多个文档之间的关系；以及

依据该预测描述符产生实体信用风险信号；

其中，所述处理器进一步用于：

为该第一组文档中包括的每个文档分配对象描述符，该对象描述符依据于历史事件或量化量度，其中，通过使日期值、实体识别符和标签值与该第一组文档中的每个文档相关联来产生该对象描述符，该日期值表示首次获得每个文档的时间段；

为该第一组文档中的每个文档分配至少一个特征向量，该至少一个特征向量是数值特征的N维向量，这些数值特征各自表示识别出来的文字内容、识别出来的语义内容，或这两种内容的组合；以及

判断多个所分配的对象描述符与多个所分配的特征向量之间的关系。

25.根据权利要求24所述的系统，其中所述处理器进一步用于：通过回归算法、支持向量机(SVM)、神经网络或决策树算法中的一项判断该关系。

26.根据权利要求24所述的系统，其中所述处理器进一步用于：依据与一个或一个以上文档相关联的元数据或指示符产生该至少一个特征向量。

27.根据权利要求24所述的系统，其中所述处理器进一步用于：通过使日期值、实体识别符和标签值与该第一组文档中的每个文档相关联来产生该对象描述符。

28.根据权利要求27所述的系统，其中所述处理器进一步用于：

判断在该第一组文档中的文档中是否识别出多个实体；以及

依据判断结果为该文档分配多个对象识别符，

其中该多个对象识别符中的每一个对象识别符对应于该多个实体中的一个实体，并且所述处理器进一步用于依据以下两项来修改该文档的标签：1)对应于该一个实体的该历史事件或该量化量度；和2)对应于该一个实体的该文档的相关度分数。

信用风险采集

技术领域

[0001] 本发明涉及风险管理，更具体来说涉及用于预测信用风险的系统和技术。

背景技术

[0002] 多年以来，风险管理一直是研究人员的一项课题，并且是商业专业人士的一个重要问题。风险管理可以帮助商业专业人士识别出一家实体(例如一家工厂)将来可能面临的不利事件，并且可以帮助确立测量、降低和管理风险的程序。如果投资人要给一些实体授信，风险管理可以帮助投资人评估这种行为可能导致的潜在损失。同样，如果投资人持有一些实体的股东权益，风险投资可以帮助投资人评估潜在的波动率(波动率会影响这类投资)，并且相应地调整他们的投资组合。

[0003] 一般来说，风险管理当中会用到许多种数据源。很多这些数据源是从包含公司在内的公开数据源直接推导而来的。例如，有许多研究人员开发出了一些信用风险模型，这些模型可以使用财会数据(例如会计比率)和定价服务机构(例如Moody's、S&P和Fitch)提供的定价数据对公司进行评级，评级依据是这些公司有多大可能拖欠债务或贷款。信用风险模型当中需要计算的量度的例子包括：拖欠债务的概率(例如，一家实体无法履行财务责任的可能性)，还有违约损失率(例如，如果发生拖欠债务的情况，那么向这家实体授信的投资人预计损失金额有多少)。

[0004] 虽然这些信息源可以给信用风险建模程序提供有价值的输入，但是这些模型忽略了大量可以公开获得的信息。例如，基于文字的数据源(例如，有些新闻文章对一家工厂的过去、目前和将来可能发生的事件所作的报道)通常会包括一些重要的信息，这些信息在信用风险建模程序中并未加以考虑。此外，这些程序通常不会分析这些数据源中包含的文字的语义语境。

[0005] 因此，需要一种改善后的信用风险建模技术，不但能够分析财会比和定价信息，而且能够分析基于文字的信息。

发明内容

[0006] 本发明公开公开了使用各种数据源来开发和实施信用风险模型的系统和技术，这些数据源包括价格数据、财会比、ESG(环境、社会和政府)数据和文字数据。每种数据源提供与工厂或公司等实体的健康状况有关的唯一并且独特的信息。这些系统和技术组合不同来源的信息，产生特别强大的信号。这些系统和技术可以用来预测许多事件，这些事件包括但不限于拖欠债务或破产的概率、违约损失率、评级机构评级变化的概率，和股本价格变动的概率。

[0007] 本发明的各种方面涉及为一组历史文档分配对象描述符和特征向量，判断这些对象描述符与特征向量之间的关系，和使用该关系为一组不同的文档分配预测描述符。

[0008] 例如，根据一个方面，判断一家公司的信用风险的方法包括：为第一组文档中包括的每个文档分配对象描述符，该对象描述符是依据历史事件或量化量度。该方法包括为该

第一组文档中的每个文档分配至少一个特征向量，该至少一个特征向量是数值特征的N维向量，这些数值特征各自表示识别出来的文字内容、识别出来的语义内容，或这两种内容的组合，以及判断多个所分配的对象描述符与多个所分配的特征向量之间的关系。该方法还包括依据这种关系为该第二组文档中不包括在该第一组文档中的每个文档分配预测描述符，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件。该方法进一步包括依据该第二组文档的至少一个预测描述符产生信号。

[0009] 在一个实施例中，该方法包括依据与文档相关联的文字内容、元数据或指示符产生至少一个特征向量。该方法可以包括：从该第一组文档中随机选择预先确定的数目的文档，形成多个文档子集，该多个子集中的每个文档包括该至少一个特征向量和相关联的第一标签值，或者不同于该第一标签值的相关联的第二标签值；计算该至少一个特征向量的每个特征的量度值，方法是通过判断该特征是描述该相关联的第一标签值还是该相关联的第二标签值；以及计算每个子集的每个特征的秩值，方法是通过汇集该多个子集中的每个特征的多个计算得来的量度值。

[0010] 该方法还可以包括：将该计算得来的秩值与预先确定的阈值进行比较；以及依据比较结果，将包括多个有阶特征和相关联的第一标签值或相关联的第二标签值的特征向量提供给机器学习模块，用于判断该关系。

[0011] 在另一个方面中，一种方法包括为第一文档分配预测描述符，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件；以及依据该预测描述符产生实体信用风险信号。在一个实施例中，该预测描述符依据于对第一训练文档与第二训练文档之间的关系的判断。

[0012] 该方法还可以包括：为该第一训练文档和该第二训练文档分配对象描述符，该对象描述符依据于历史事件或量化量度；以及为该第一训练文档和该第二训练文档分配至少一个特征向量，该至少一个特征向量是识别出来的文字内容、识别出来的语义内容，或这两种内容的组合的数值表示。

[0013] 在另一方面中，一种装置包括：用于为第一文档分配预测描述符的装置，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件。该装置还包括用于依据该预测描述符产生实体信用风险信号的装置。

[0014] 在一个实施例中，该预测描述符是依据第一训练文档，该第一训练文档具有对应的第一对象描述符和对应的第一存储的数值表示，该对应的第一对象描述符和该对应的第一存储的数值表示具有第一预定义关系。该预测描述符也可以依据第二训练文档。该第二训练文档包括对应的第二对象描述符和对应的第二存储的数值表示，这两者具有第二预定义关系。

[0015] 在又一实施例中，一种系统包括：数据存储器，包括第一组文档和第二组文档；以及服务器，包括处理器和存储器，该存储器存储响应于接收到访问服务的请求而使得该处理器进行以下操作的指令：为第二组文档中的第一文档分配预测描述符，该预测描述符表示与公司拖欠债务、公司破产、违约损失率、信贷息差、评级机构评级变化和股本价格变动中的至少一项有关的将来事件。该预测描述符依据于第一组文档中的多个文档之间的关

系。该系统的处理器依据该预测描述符产生实体信用风险信号。

[0016] 本发明还公开了另外的系统、方法以及包括存储有用于实施各种技术的机器可读指令的机器可读媒体的物品。下文更详细地论述各种实施方案的细节。

附图说明

[0017] 图1是示范性的基于计算机的信用风险采集系统的示意图。

[0018] 图2图解说明根据本发明的一个实施例的用于判断一家实体的信用风险的示范性方法。

[0019] 图3图解说明根据本发明的一个实施例的用于为机器学习算法提供与文档相关联的有阶特征向量的示范性方法。

[0020] 各图中的相同参考符号表示相同元件。

具体实施方式

[0021] 图1所示为用于分析与一家实体(例如,一家公司或工厂)相关联的信用风险的基于计算机的系统10。系统10经配置以使用一个或一个以上文本挖掘组件来分析文档中所包含的财会比、定价数据,以及基于文字的数据。有利的一点是,系统10可以用来预测与一家实体有关的许多项数据,例如拖欠债务或破产的概率,违约损失率,评级机构(例如,S&P、Moody's、Fitch)的评级发生变化的概率,以及股本价格大幅变动的概率。

[0022] 如图1的实例中所示,系统10包括服务器装置12,所述服务器装置12经配置以包括处理器14,例如中央处理单元(CPU),随机存取存储器(RAM)16,一个或一个以上输入—输出装置18,例如显示器装置(未图示),以及键盘(未图示),还有非易失性存储器20,所有这些元件都通过公共总线22相互连接,而且由处理器14加以控制。在一个实施例中,如图1所示,非易失性存储器20经配置以包括:文档模块24,用于给文档分配对象描述符,每个对象描述符包括一个文档标签;特征模块26,用于从加了标签的文档中提取特征向量;选择模块28,用于根据相关度给特征向量排序;学习模块30,用于判断多个对象描述符与分配的特征向量之间的关系;以及预测模块31,用于根据分配给一个文档的至少一个预测描述符来产生信号。所述预测描述符表示一家实体的将来事件,并且是根据文档中的识别出的文字内容、文档中的识别出的文字的语义语境,或者是这两项的组合。下文更具体地论述文档模块24、特征模块26、选择模块28、学习模块30和预测模块31(统称为"文本挖掘模块")的其它细节。

[0023] 系统10可以经配置以包括访问装置(未图示),该装置通过网络与服务器装置12通信。所述访问装置可以包括个人计算机、膝上型计算机,或者其它类型的电子装置,例如蜂窝电话或个人数字助理(PDA)。例如,在一个实施例中,访问装置连接到I/O装置(未图示),该I/O装置包括键盘,与鼠标等指向装置配合,用于向服务器12发送请求。优选的情况是,访问装置的存储器经配置以包括浏览器,该浏览器用于通过网络从服务器12请求并接收信息。

[0024] 该网络可以包括各种装置,例如路由器、服务器和交换元件,这些装置用内联网、互联网或因特网配置连接在一起。在一些实施方案中,该网络使用有线通信在访问装置与服务器装置12之间传送信息。在另一个实施例中,该网络使用无线通信协议。在其它实施例中,该网络使用有线技术与无线技术的组合。

[0025] 文本挖掘模块24、26、28、30、31适于处理各种来源提供的数据。例如，在一个实施例中，一个或一个以上文本挖掘模块24、26、28、30访问和处理汤姆森路透集团(Thomson Reuters)提供的文字数据的独家且及时的来源，例如汤姆森路透街头事件(Thomson Reuters StreetEvents)数据馈送提供的电话会议记录，还有路透社新闻视野档案(Reuters NewsScope Archive)的新闻文章和头条。这一个或一个以上文本挖掘模块24、26、28、30、31还可以访问和处理一些公司的年度和季度财务报表和股票经纪人研究文档。

[0026] 在一个实施例中，文本挖掘模块24、26、28、30、31被配置在服务器装置12的非易失性存储器20中，并且在处理文字数据时实施一项“词语包”技术。例如，在一个实施例中，文档模块24将文字数据当做词语和短语的集合来处理，相对来说不太注意词语或短语在文档中的位置，或文字的语法和语言特性。在另一实施例中，文档模块24经配置以访问ClearForest等第三方机构根据文档文字推导而来的特征向量和指示符。推导得来的指示符是依据接受分析的文字的语言特性，并且为文档文字提供语义语境。

[0027] 文档数据存储器32是一个存储装置，一个或一个以上文本挖掘模块24、26、28、30、31利用这个存储装置来访问和存储与一个或一个以上接收到的文档40、一组或一组以上文档34、与一个或一个以上文档相关联的特征向量和标签38，以及学习模块30使用的训练数据38有关的信息。在一个实施例中，文档数据存储器32是关系数据库。在另一实施例中，文档数据存储器32是目录服务器，例如轻型目录访问协议(LDAP)服务器。在其它实施例中，数据存储器32是在装置服务器12的非易失性存储器20中配置的一个区域。虽然图1所示的数据存储器32连接到服务器12，但是本领域的技术人员明白，文档数据存储器32可以分布在各种服务器上，并且可以由服务器装置12访问。

[0028] 文档模块24将基于文字的历史文档组织成一组或一组以上文字数据。例如，在一个实施例中，文档模块24组织和分析电话会议记录，新闻文章，财务报表，券商研究出版物，在线公开内容(例如，博客和推特)，还有其它基于文字的数据源。文档模块24将每份基于文字的文档当做独特的观察结果来处理。在一个实施例中，文档模块24将多份文档汇集在一起，然后把这些文档当成单独一份文档加以处理。例如，在一个实施例中，文档模块24将某一段时间(例如，一天，一周、一个月等等)与一家实体有关的所有文档都当做一份文档加以处理。本文中使用的词语“文档”包括单独一份文档和/或一堆文档和/或一份文档的一部分，或者是以上内容的组合。文档模块24还可经配置以仅仅处理文档的特定章节，或者对文档的某些部分做不同处理。例如，在一个实施例中，文档模块24对电话会议记录的管理讨论部分的分析与对电话会议记录的问答部分的分析有所不同。

[0029] 文档模块24给每个文档分配一个“截止日期”，用来表示可以获得文档的日期。在一个实施例中，这个日期是这份文档首次通过提供文档的数据馈送或数据传递平台可被用户或客户访问的日期。此外，文档模块24将每份文档与一个或一个以上公司相关联。在一个实施例中，文档模块24根据随每份文档提供的元数据来判断与公司的关联情况。在另一实施例中，文档模块24搜索文档的文字内容，寻找文档中包括的实体名称、公司股票行情或者实体识别符，然后将这些搜索结果信息与预先识别的实体识别符进行比较。

[0030] 文档模块24使用公司与日期分配情况来构造一个或一个以上标签文档训练集合。文档模块24将一组文档中的每份文档标记为“阳性”(例如，表示该事件)或者“阴性”(例如，并未表示该事件)。例如，在一个实施例中，在设法预测从提供文档之日起12个月内一家实

体破产和拖欠债务的可能性时，如果与该文档相关联的实体自文档“截止日期”起十二(12)个月内拖欠债务或者破产，则文档模块24将一个集合中的每个前期接收到的文档的文档标签值都设置成“阳性”值。如果与该文档相关联的实体自文档“截止日期”起12个月内未拖欠债务和破产，则文档模块24将该文档的标签值设置成“阴性”值。

[0031] 如果目标是要预测多项数据，则文档模块24可以为每份文档设置多个标签。例如，一个标签可以表示将来破产或拖欠债务，第二个标签可以表示将来股本价格大幅下降。或者，在另一个实施例中，文档模块24将单个标签设置成预测多个事件。例如，在一个实施例中，文档模块24将一个标签设置成表示将来破产或拖欠债务和将来股本价格大幅下降两项内容。

[0032] 例如，在一个实施例中，文档模块24为每份文档产生和分配一个对象描述符，方法是将前面提到的“截止日期”、实体识别符和标签值与每份文档相关联。在另一个实施例中，文档模块24根据量化量度给每份文档分配一个对象描述符。例如，这个量化量度的根据可以是感情评级、语气评级、公司事件指示符、语言分析指示符、统计分析指示符，或者以上内容的组合。在一个实施例中，语言分析指示符与文档中包括的词语(例如，名词短语)的结构或语法有关。在另一个实施例中，统计分析指示符与该文档中一些词语的出现频率有关。

[0033] 在一些例子中，一份文档可能涉及不止一家实体。例如，一篇技术类新闻文章中可能探讨了微软与苹果电脑两家公司。当发生这种情况时，文档模块24经配置以使用根据文档文字推导得来的语义和统计信息，识别出这份文档与这两家公司的相关度。例如，如果前面提到的这篇技术类新闻文章中主要是探讨微软公司发布一种新型操作系统，然后简短地介绍了这种新型操作系统可能会如何减少苹果的市场份额，那么，文档模块24判断，这篇文章中有百分之七十(70％)与微软公司有关，百分之三十(30％)与苹果电脑公司有关。在一个实施例中，文档模块24针对文章中识别出的每家实体为该文档分配一个对象描述符。文档模块24将对象描述符的日期值设置成一个常数值，并且将实体标识符设置成对应于文档中识别出的每家公司。此外，在一个实施例中，文档模块24经配置以根据该文档与相应实体的相关度来修改文档的标签。

[0034] 特征模块26将每份文档表示成一组特征向量。本文中使用的短语“特征向量”是指数值特征的N维向量或数组，其中每个数值特征表示每份文档中包括的识别出的文字内容、识别出的语义内容，或以上内容的组合。这些特征向量是根据文档本身的文字推导得来的。特征模块26可以根据文档产生一些特征向量，这些文档包括根据文字推导得来的元数据或指示符。例如，可以给文档分配“感情”评级(正面，中立，或负面)、“语气”评级(正面，负面，或中立)，表示对合并或收购等公司事件的讨论的指示符，或者根据整个文档文字的语言或数据分析推导得来的其它类型的指示符，上文已结合文档模块24对这些操作做了说明。这些指示符可以由ClearForest等第三方机构产生，也可以由特征模块26自己产生。特征模块26还可以经配置以根据每份文档中的个别词语和短语来产生特征。

[0035] 例如，在一个实施例中，特征模块26将一份文档分成一些词语和长度为几个词的短语(k-mers/n-grams)。特征模块26提取所有连续的词语，和长度为三个或不到三个词语的短语。然后，特征模块26去掉标点符号和纯粹的数值串。

[0036] 在一个实施例中，特征模块26使用“无用词列表”去掉被视为不相关的特定词语。常见的无用词包括“这个”、“和”、“一个”等等词语。特征模块26还可经配置以根据适当的名

词、人物、公司或实体名称、公司或行业特有的术语、存在于文档数据存储器中的词语的统计特性的列表或者由分析人员或专家提供的定制列表来使用其它无用词。特征模块26还经配置以去掉可能与事件或者时间段特有的因素相关的无关词语或非鲁棒性词语。

[0037] 在一个实施例中，特征模块26对词语长度提出要求，包括长度在两个(2)与二十个(20)字符之间的词语，并且去掉其它所有词语。有利的一点是，这种做法有种效果，就是能把特征向量限制到典型的词语和短语，并且去掉混乱的文字。特征模块26还可以经配置为文档中的所有词语提取词干(这个过程可以将词形经过变化的词语和/或衍生的词语还原成词干、基本形态或词根形式)。例如，在一个实施例中，特征模块26使用提取词干算法(例如，本领域已知的波特提取词干算法(Porter stemmer))来提取文档中的词语的词干，去掉词语中的时态、词形变化和复数形式。

[0038] 一旦特征模块26产生了特征值，选择模块28就可以判断哪些特征最有助于进行预测。选择模块28使用下文所述的一个或一个以上统计量度或测试方法独立地评估每个特征的预测能力。这样，选择模块28可以消除噪声或无关特征。选择模块28还可以经配置以消除非常罕见的特征，也就是只会在非常少数的文档或者在文档的一小部分中出现的特征。

[0039] 选择模块28可以选出信息量最大的特征，方法是重复检查可以获得的文档的不同子集34。选择模块28从所有可以获得的文档的集合34中随机选出一个文档子集，从一组标记为阳性的文档中选出一些文档，和从一组标记为阴性的文档的中选出一些文档。在一个实施例中，这些文档子集是平衡的，也就是说，每个子集中阳性文档与阴性文档的数目相同。或者，一个或一个以上子集可以是不平衡的，也就是阳性文档与阴性文档的数目不同。本文中的这个文档子集也称为“文集”。在一个实施例中，选择模块28选择平衡的文集，其中含有大概三分之一(1/3)的阳性文档，和相等数目的阴性文档。

[0040] 如上所述，选择模块28评估文集中的每个特征的预测能力。选择模块28首先去掉文集中的最少数目的文档中未出现的所有特征。例如，在一个实施例中，选择模块28去掉文集文档中的至少百分之一(1％)中未出现的所有特征。在此过程中，选择模块28会去掉所有非常罕见的特征，这些通常都是噪声和不可靠的指示符。罕见特征的实例包括各个名和姓。然后，选择模块28会使用一个或一个以上统计量度或测试(例如准确度、检索率、精确性、信息增益、两项分离或其它量度)来评估每个剩余特征。这些量度可以测定该特征在文集中的被标记为阳性的文档和被标记为阴性的文档之间加以区分的能力。

[0041] 选择模块28重复上述过程若干次，从而重复地产生一个随机选择的文集，并且测量该文集中各特征的表现能力。一般来说，如果整组文档非常不平衡(例如，如果该组中标记为阳性的文档的数目远远多于标记为阴性的文档)，则选择模块28会产生非常多的文集。选择模块28使用文档的许多小型子集对整个文档集合进行采样。例如，在一个实施例中，建立了五十(50)到两百个(200)文集。因为重复地产生随机选择的文集，所以可能多个文集选择整组文档中的第一文档，而任何文集都不选择整组文档中的第二文档。

[0042] 在为每个文集中的特征记分后，选择模块28为所有文集中的每个特征计算总体表现能力。在一个实施例中，选择模块28计算每个特征的文集分数的平均值。然后，选择模块28选择总体分数最高的特征当做最终特征。在一个实施例中，选择模块28要求这个特征在最少的文集中有最高的得分(例如，这个特征在该文集的百分之二十(20％)中的前一千个(1000)特征当中)，以便确保这个特征通常表现出色。选择模块28包括的最终特征的数目可

以设置成特定的截止值。例如，在一个实施例中，选择模块28选择满足所有上述标准的总体分数最高的前两百五十个(250)特征。有利的情况是，这个截止值可以依据最终表现的鲁棒性，并且可以由受到过特征向量相关训练的机器学习算法的预测能力来判断，对此下文有更具体地说明。选择模块28经配置以选择尽可能少的特征，同时尽可能提高检测破产和拖欠债务或其它事件的能力。

[0043] 此外，选择模块28可以对特征应用丛集算法(例如，层级丛集或k均值丛集)或降维算法(例如，主成分分析(PCA)或非负矩阵分解(NMF))，将可获得的特征的数目减少或压缩为额外特征的集合。选择模块28可以在上述过程之前、上述过程之后或独立于上述过程应用这些技术。

[0044] 选择模块28将特征向量中的最终特征提供给学习模块30，每个向量中的一个元素对应于一个最终特征。该特征的数值表示可以是二进制值(例如，如果文档中存在特定特征，则值＝“1”，或者，如果没有特定特征，则值＝“0”)，或者是依据序数(例如，特征在文档中的出现次数)，或者是依据出现频率(例如，特征在文档中的出现次数通过独特的数字来归一化，例如文档的数目、特征的数目等等)。

[0045] 一旦选择模块28提供了所有可获得的文档的特征向量，那么，不论在特征选择过程中文集中是否包括某一个文档，学习模块30都会判断特征向量与其相关对象描述符之间的关系。学习模块30使用这种关系来预测以前未看到的文档的文档标签(例如，预测描述符)。学习模块30判断这种关系的方法是：将特征向量提供到统计预测或机器学习算法中，例如回归算法、支持向量机(SVM)、神经网络或决策树算法。在一个实施例中，学习模块30实施一种或一种以上“boosting”或“bagging”技术，这些技术在本领域中是已知的。学习模块30还可经配置以使用综合技术和/或组合几种不同的机器学习算法。例如，在一个实施例中，学习模块30通过对文档特征向量子集使用许多较小SVM，判断这些关系，并借此受到训练，然后允许每个SVM独立地为文档标签投票。在一个实施例中，学习模块30使用交叉验证等模型调谐技术，判断要使用的特征的最佳数目，并且对预测模型进行调谐。

[0046] 一旦调谐了学习模块30，预测模块31就对新文档应用这个学习模块，预测新文档的标签。预测可以是一个概率，或者是一个连续变量(例如，被分配给每个新文档的预测描述符)。例如，在一个实施例中，预测描述符是零(0)到一(1)的数字，其中零(0)表示发生拖欠债务或破产的可能性很小，且一(1)表示发生拖欠债务或破产的可能性很大。

[0047] 预测模块31使用预测描述符产生公司特有的信号。在一个实施例中，所述公司信号是依据与一个或一个以上新近接收到的文档相关联的预测描述符。在另一实施例中，所述公司信号是依据某个时间段(一天、一周、一个月等等)接收到的文档中的预测描述符的汇总。所述预测描述符可能要求进行变换或校准，以便使拖欠债务事件的历史或基线水平正确地对准。

[0048] 图2中公开了根据本发明的一个实施例判断一家公司的信用风险的示范性方法。如图2的实例中所示，在一个实施例中，文档模块24首先产生第一文档集合中包括的每个文档的对象描述符，方法是通过使一个日期值、一个实体识别符和一个标签值与每个文档相关联(50)，然后将每个产生出的对象描述符分配给第一文档集合中包括的每个文档(52)。接下来，特征模块26依据与每个文档相关联的元数据或指示符为第一文档集合中包括的每个文档产生至少一个特征向量(54)，然后将产生出的每个特征向量分配给第一文档集合中

的每个文档(56)。然后,学习模块30判断多个所分配的对象描述符与多个所分配的特征向量之间的关系(58)。下文将说明图2的功能步骤60、62和64。

[0049] 现在参照图3,图中公开了在判断要将哪些特征向量提供给学习模块30时选择模块28执行的示范性步骤。如图3的实例中所示,首先,选择模块28从第一文档集合中随机选择预先确定的数目的文档,形成多个文档子集(70)。接下来,选择模块28计算这多个子集中的每个特征的量度值,方法是判断这个特征是描述与文档相关联的第一标签值还是第二标签值(72)。然后,选择模块28计算每个子集的每个特征的秩值,方法是汇集多个子集中的每个特征的多个计算得来的量度值(74)。

[0050] 一旦计算出秩值,选择模块28就将计算出来的秩值与一个或一个以上预先确定的阈值比较(76),并且根据比较结果将具有相关联的第一标签值或第二标签值的多个有阶特征提供给学习模块30(78)。在一个实施例中,如果一个秩值满足或超过了该阈值,则将这个特征包括在被提供给学习模块30的特征向量中。

[0051] 回头参照图2,一旦学习模块30使用有阶特征值判断了这个关系,预测模块31便使用由学习模块30判断的关系,产生将与第二文档集合中不包括在第一文档集合中的每个文档相关联的预测描述符(60)。然后,预测模块31将每个产生出的预测描述符分配给第二文档集合中的每个文档(62),并且依据第二文档集合的至少一个预测描述符产生一个信号(64)。

[0052] 该系统可以经配置以预测商业部门内的破产或拖欠债务。例如,在一个实施例中,该系统经配置以依据与一个文档相关联的其它实体的性质或特性来识别一家实体。例如,探讨微软公司的最新操作系统的一份文档可能与信息技术(IT)部门相关联,因为在这份文档中识别出来微软公司,而微软公司通常归类为IT部门。在一个实施例中,该系统使用部门识别符作为实体识别符来构造对象描述符,并且将标签值设置为接下来十二个(12)月该部门存在或是不存在破产或拖欠债务情况。

[0053] 该系统的各种特征可以在硬件、软件或硬件与软件的组合中实施。例如,该系统的一些特征可以在可编程计算机上运行的一个或一个以上计算机程序中实施。每个程序可以在高级程序语言或面向对象的编程语言中实施,以便与计算机系统或其它机器通信。此外,每个这种计算机程序可以存储在可由通用或专用可编程计算机或处理器读取的只读存储器(ROM)等存储媒体上,用于配置和操作计算机来执行上述功能。

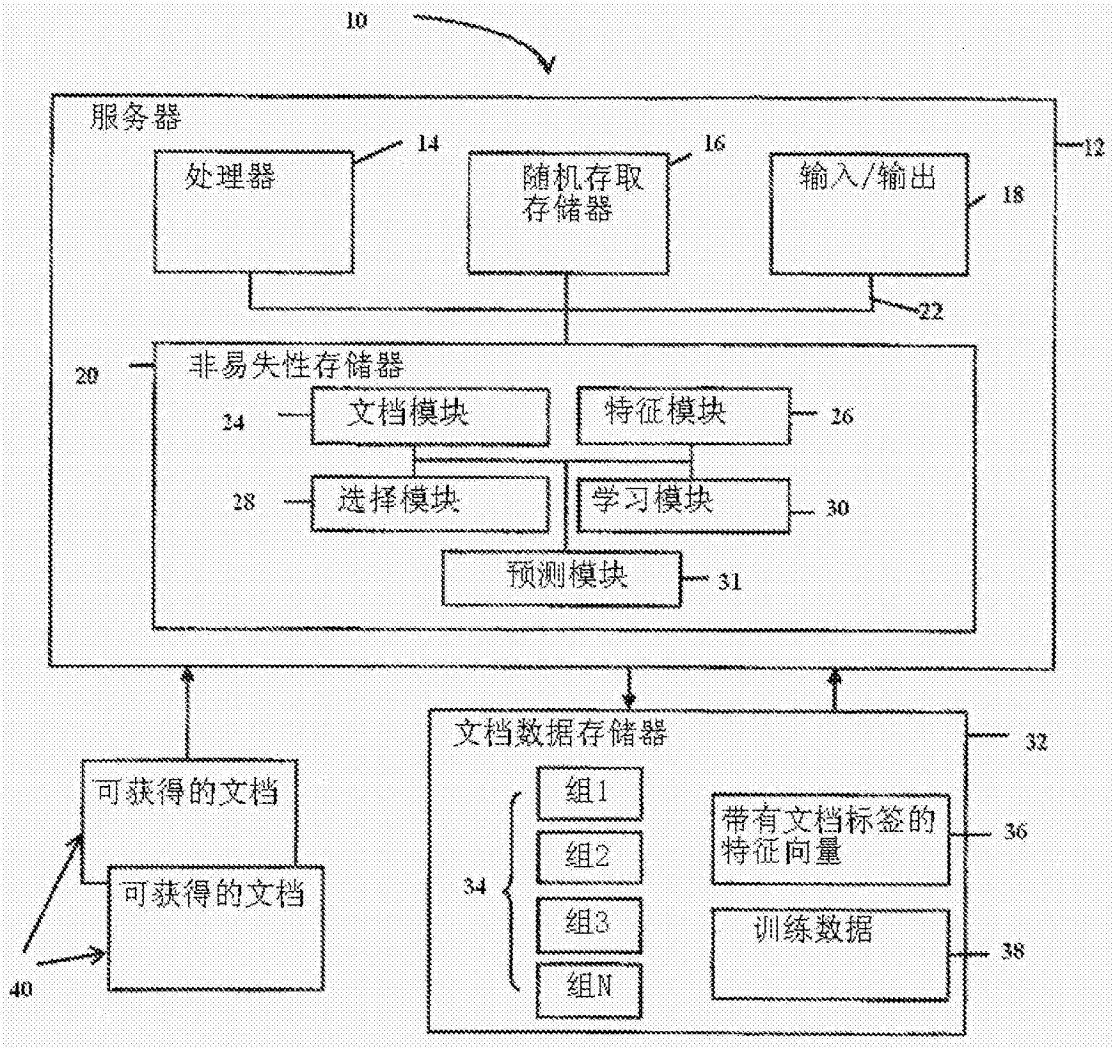

10
服务器
12
处理器
14
随机存取存储器
16
输入/输出
18
22
非易失性存储器
20
文档模块
24
特征模块
26
选择模块
28
学习模块
30
预测模块
31
文档数据存储器
32
组1
组2
组3
组N
34
带有文档标签的特征向量
36
训练数据
38
可获得的文档
可获得的文档
40

图1

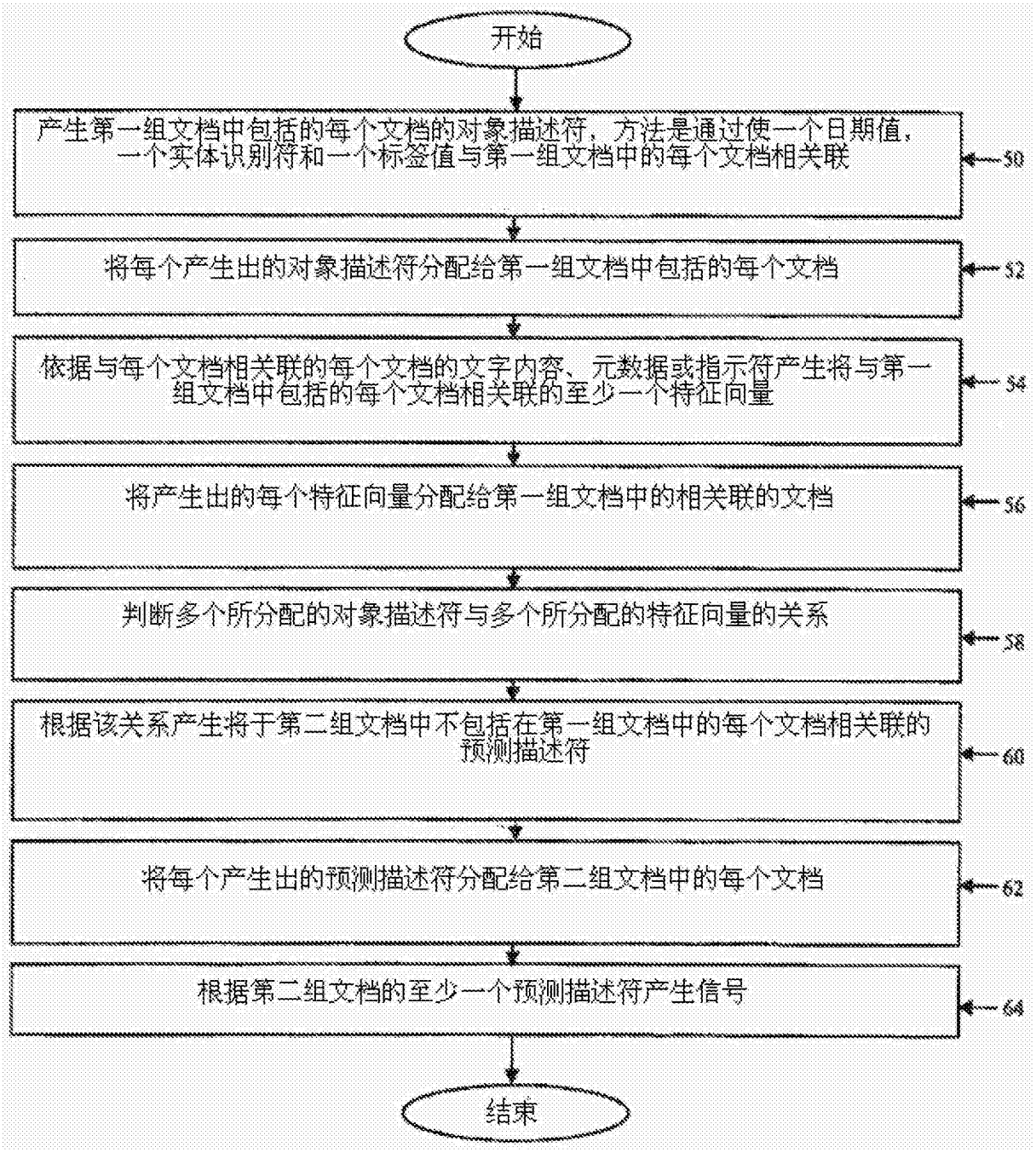

开始
产生第一组文档中包括的每个文档的对象描述符，方法是通过使一个日期值，一个实体识别符和一个标签值与第一组文档中的每个文档相关联
50
将每个产生出的对象描述符分配给第一组文档中包括的每个文档
52
依据与每个文档相关联的每个文档的文字内容、元数据或指示符产生将与第一组文档中包括的每个文档相关联的至少一个特征向量
54
将产生出的每个特征向量分配给第一组文档中的相关联的文档
56
判断多个所分配的对象描述符与多个所分配的特征向量的关系
58
根据该关系产生将于第二组文档中不包括在第一组文档中的每个文档相关联的预测描述符
60
将每个产生出的预测描述符分配给第二组文档中的每个文档
62
根据第二组文档的至少一个预测描述符产生信号
64
结束

图2

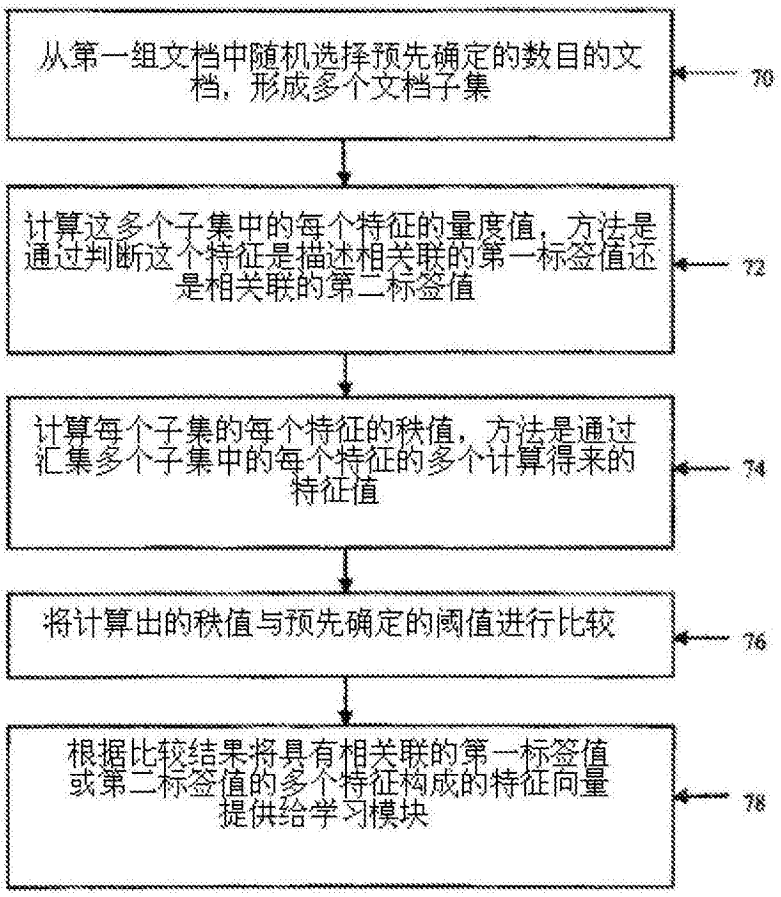

从第一组文档中随机选择预先确定的数目的文档，形成多个文档子集
70
计算这多个子集中的每个特征的量度值，方法是通过判断这个特征是描述相关联的第一标签值还是相关联的第二标签值
72
计算每个子集的每个特征的秩值，方法是通过汇集多个子集中的每个特征的多个计算得来的特征值
74
将计算出的秩值与预先确定的阈值进行比较
76
根据比较结果将具有相关联的第一标签值或第二标签值的多个特征构成的特征向量提供给学习模块
78

图3