(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104794209 A
(43)申请公布日 2015.07.22

(21)申请号 201510201444.6

(22)申请日 2015.04.24

(71)申请人 清华大学
地址 100084 北京市海淀区 100084-82 信箱
申请人 清华大学无锡应用技术研究院

(72)发明人 徐华

(74)专利代理机构 北京清亦华知识产权代理事务所（普通合伙）11201
代理人 张大威

(51)Int.Cl.
G06F 17/30(2006.01)

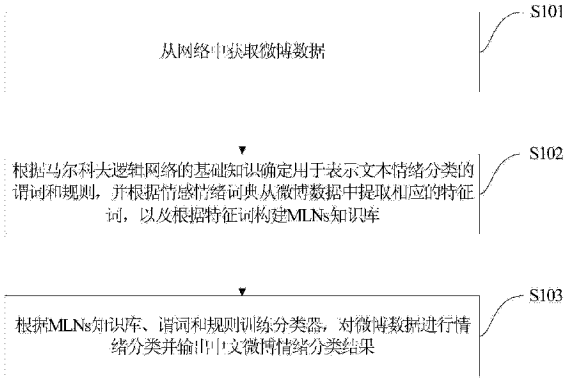
权利要求书2页 说明书6页 附图2页

(54)发明名称

基于马尔科夫逻辑网络的中文微博情绪分类方法及系统

(57)摘要

本发明提供一种基于马尔科夫逻辑网络的中文微博情绪分类方法及系统，该方法包括：从网络中获取微博数据；根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从微博数据中提取相应的特征词，以及根据特征词构建MLNs知识库；根据MLNs知识库、谓词和规则训练分类器，对微博数据进行情绪分类并输出中文微博情绪分类结果。本发明的实施例可以准确地对微博进行情绪分类。



CN 104794209 A

1. 一种基于马尔科夫逻辑网络的中文微博情绪分类方法，其特征在于，包括以下步骤：

从网络中获取微博数据；

根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从所述微博数据中提取相应的特征词，以及根据所述特征词构建 MLNs 知识库；

根据所述 MLNs 知识库、所述谓词和规则训练分类器，对所述微博数据进行情绪分类并输出中文微博情绪分类结果。

2. 根据权利要求 1 所述的基于马尔科夫逻辑网络的中文微博情绪分类方法，其特征在于，所述从网络中获取微博数据，进一步包括：利用网络爬虫充微博页面中抓取所述微博数据。

3. 根据权利要求 1 所述的基于马尔科夫逻辑网络的中文微博情绪分类方法，其特征在于，所述谓词和规则表示所述微博数据以及微博数据之间的关联关系。

4. 根据权利要求 1 所述的基于马尔科夫逻辑网络的中文微博情绪分类方法，其特征在于，所述微博数据包括训练数据集和测试数据集。

5. 根据权利要求 4 所述的基于马尔科夫逻辑网络的中文微博情绪分类方法，其特征在于，所述根据所述 MLNs 知识库、所述谓词和规则训练分类器，进一步包括：

根据所述训练数据集训练所述分类器，其中，在训练所述分类器时，根据证据谓词和查询谓词进行规则参数的学习；

根据所述测试数据集测试所述分类器的性能，其中，在测试所述分类器的性能时，采用准确率、召回率和 F 值作为所述分类器的性能评估指标。

6. 一种基于马尔科夫逻辑网络的中文微博情绪分类系统，其特征在于，包括：

数据获取模块，用于从网络中获取微博数据；

马尔科夫逻辑表示模块，用于根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从所述微博数据中提取相应的特征词，以及根据所述特征词构建 MLNs 知识库；

分类器训练模块，用于根据所述 MLNs 知识库、所述谓词和规则训练分类器，对所述微博数据进行情绪分类并输出中文微博情绪分类结果。

7. 根据权利要求 6 所述的基于马尔科夫逻辑网络的中文微博情绪分类系统，其特征在于，所述数据获取模块用于利用网络爬虫充微博页面中抓取所述微博数据。

8. 根据权利要求 6 所述的基于马尔科夫逻辑网络的中文微博情绪分类系统，其特征在于，所述谓词和规则表示所述微博数据以及微博数据之间的关联关系。

9. 根据权利要求 6 所述的基于马尔科夫逻辑网络的中文微博情绪分类系统，其特征在于，所述微博数据包括训练数据集和测试数据集。

10. 根据权利要求 9 所述的基于马尔科夫逻辑网络的中文微博情绪分类系统，其特征在于，所述分类器训练模块用于：

根据所述训练数据集训练所述分类器，其中，在训练所述分类器时，根据证据谓词和查询谓词进行规则参数的学习；

根据所述测试数据集测试所述分类器的性能，其中，在测试所述分类器的性能时，采用

准确率、召回率和 F 值作为所述分类器的性能评估指标。

基于马尔科夫逻辑网络的中文微博情绪分类方法及系统

技术领域

[0001] 本发明涉及计算机应用技术与互联网技术领域，特别涉及一种基于马尔科夫逻辑网络的中文微博情绪分类方法及系统。

背景技术

[0002] 随着互联网时代的快速发展及社交网络的迅速升起，越来越多的人开始使用微博来发表自己的言论，它已经成为人们生活工作中的一个大众舆论平台。中国互联网络信息中心的调查研究报告显示，2014 年 4 月 20 日到 2014 年 5 月 10 日，40.7％的用户每天都会登陆微博，25.9％的用户每周会登陆 2 次以上，用户活跃度和用户黏性均比较高。用户可以在微博上自由地发表自己的见解、表达自己的情绪。鉴于微博的开放性、大众性、交互性强、传播速度快等特征，使得微博信息直接影响着国家的民生、经济及社会稳定，分析其中蕴含的巨大的价值信息，有助于便捷地传播新闻资讯、直接展示个人意见、推动公益事业的发展、推动政府政务透明化等等。对于文本的情绪观点解析，一直是自然语言处理领域的热点。先前对于文本情绪观点挖掘及分类已经做了许多研究，有基于机器学习的 SVM 和 CRF 分类方法，也有基于语言规则的分类方法，还有利用深度学习方法的。但是这些都是对每个文本单独分类，忽略了类别之间的联系，分类结果准确度较低。

发明内容

[0003] 本发明的目的旨在至少解决上述的技术缺陷之一。

[0004] 为此，本发明的目的在于提出一种基于马尔科夫逻辑网络的中文微博情绪分类方法。该方法可以准确地对微博进行情绪分类。

[0005] 本发明的另一个目的在于提出一种基于马尔科夫逻辑网络的中文微博情绪分类系统。

[0006] 为了实现上述目的，本发明的第一方面的实施例公开了一种基于马尔科夫逻辑网络的中文微博情绪分类方法，包括以下步骤：从网络中获取微博数据；根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从所述微博数据中提取相应的特征词，以及根据所述特征词构建 MLNs(Markov Logic Networks，MLNs) 知识库；根据所述 MLNs 知识库、所述谓词和规则训练分类器，对所述微博数据进行情绪分类并输出中文微博情绪分类结果。

[0007] 根据本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法，通过马尔科夫逻辑网络，利用其参数学习和推理体系，对微博文本进行建模，然后将其分成例如高兴、悲伤、厌恶、愤怒、恐惧、惊奇等六类情绪类别，具有分类准确的优点。

[0008] 另外，根据本发明上述实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法还可以具有如下附加的技术特征：

[0009] 在一些示例中，所述从网络中获取微博数据，进一步包括：利用网络爬虫充微博页面中抓取所述微博数据。

[0010] 在一些示例中，所述谓词和规则表示所述微博数据以及微博数据之间的关联关系。

[0011] 在一些示例中，所述微博数据包括训练数据集和测试数据集。

[0012] 在一些示例中，所述根据所述 MLNs 知识库、所述谓词和规则训练分类器，进一步包括：根据所述训练数据集训练所述分类器，其中，在训练所述分类器时，根据证据谓词和查询谓词进行规则参数的学习；根据所述测试数据集测试所述分类器的性能，其中，在测试所述分类器的性能时，采用准确率、召回率和 F 值作为所述分类器的性能评估指标。

[0013] 本发明第二方面的实施例公开了一种基于马尔科夫逻辑网络的中文微博情绪分类系统，包括：数据获取模块，用于从网络中获取微博数据；马尔科夫逻辑表示模块，用于根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从所述微博数据中提取相应的特征词，以及根据所述特征词构建 MLNs 知识库；分类器训练模块，用于根据所述 MLNs 知识库、所述谓词和规则训练分类器，对所述微博数据进行情绪分类并输出中文微博情绪分类结果。

[0014] 根据本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类系统，通过马尔科夫逻辑网络，利用其参数学习和推理体系，对微博文本进行建模，然后将其分成例如高兴、悲伤、厌恶、愤怒、恐惧、惊奇等六类情绪类别，具有分类准确的优点。

[0015] 另外，根据本发明上述实施例的基于马尔科夫逻辑网络的中文微博情绪分类系统还可以具有如下附加的技术特征：

[0016] 在一些示例中，所述数据获取模块用于利用网络爬虫充微博页面中抓取所述微博数据。

[0017] 在一些示例中，所述谓词和规则表示所述微博数据以及微博数据之间的关联关系。

[0018] 在一些示例中，所述微博数据包括训练数据集和测试数据集。

[0019] 在一些示例中，所述分类器训练模块用于：根据所述训练数据集训练所述分类器，其中，在训练所述分类器时，根据证据谓词和查询谓词进行规则参数的学习；根据所述测试数据集测试所述分类器的性能，其中，在测试所述分类器的性能时，采用准确率、召回率和 F 值作为所述分类器的性能评估指标

[0020] 本发明附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0021] 本发明上述的和 / 或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中，

[0022] 图 1 是根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法的总体流程图；

[0023] 图 2 是根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法的实施步骤图；以及

[0024] 图 3 是根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类系统的结构框图。

具体实施方式

[0025] 下面详细描述本发明的实施例，实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

[0026] 在本发明的描述中，需要理解的是，术语“中心”、“纵向”、“横向”、“上”、“下”、“前”、“后”、“左”、“右”、“竖直”、“水平”、“顶”、“底”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0027] 在本发明的描述中，需要说明的是，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0028] 在描述本发明实施例的方法之前，首先对马尔科夫逻辑表示方法以及分类器训练方法进行说明。

[0029] 用马尔科夫逻辑网方法来做情绪分类，采取的主要特征为词，也就是某条微博文本中的某些词。马尔科夫逻辑表示主要包括三大部分：谓词、规则和知识库。谓词由一个谓词声明列表组成，每一个谓词声明都指定一个带有参数类型列表的谓词名字，每一个类型都由常量集支持，该常量集由规则和证据中相应的谓词参数填充。据此，可以将文本分类用三个谓词表示，如表 1 所示。

[0030] 表 1

[0031]

谓词	含义
hasWord(word, text)	表示文本 text 中有词 word
topic(class, text)	表示文本 text 属于 class 情绪类别
sameCat(class1, class2)	表示情绪类别 class1 与情绪类别 class2 相同

[0032] 表 1 示出了基于马尔科夫逻辑网络的中文微博情绪分类谓词列表。

[0033] MLNs 的规则由谓词及连接词、量词、运算符等联结而成。根据上面的三个谓词，可以得到以下两个公式：

[0034] hasWord(w,t1),hasWord(w,t2), ! topic(c,t1) =>! topic(c,t2) （公式 1）

[0035] ! sameCat(c1,c2),topic(c1,t) =>! topic(c2,t) （公式 2）

[0036] 其中，公式 1 表示如果文本 t1 中有词 w，文本 t2 中也有词 w，且文本 t1 不属于类别 c，那么文本 t2 也不属于类别 c；公式 2 表示如果类别 c1 与类别 c2 不相同，且文本 t 属于类别 c1，那么该文本不属于类别 c2。并且赋予公式 1、公式 2 的权重数值分别为 1、5。

[0037] 知识库由一个或多个证据或查询文件构成。每个文件都由一个原子列表组成。本发明所使用的知识库是根据文本中的关键词及谓词方案构建而成的。

[0038] 马尔可夫逻辑网络是最近发展起来的、强大的关于复杂依赖关系数据的学习和推理的形式体系。从概率的角度看，马尔科夫逻辑网提供了一种简洁的语言来定义大型的马尔科夫网，能够灵活地、模板化地与大量知识相结合；从一阶逻辑的角度看，马尔科夫逻辑网能够健全地处理不确定性、容错性，甚至矛盾的知识库，降低了脆弱性。

[0039] MLNs 的权重学习是在给定一个训练数据集和 MLNs 程序的前提下，进一步学习和优化 MLNs 规则的最优权重，从而最大化训练数据集的可能性。MLNs 的推理即在给定学习的公式及测试数据集证据谓词的情况，正确预测查询谓词。

[0040] 马尔科夫逻辑网是一种结合概率与一阶逻辑的统计学习方法。从概率的角度看，马尔科夫逻辑网提供了一种简洁的语言来定义大型的马尔科夫网，能够灵活地、模板化地与大量知识相结合；从一阶逻辑的角度看，马尔科夫逻辑网能够健全地处理不确定性、容错性，甚至矛盾的知识库，降低了脆弱性。正是由于其的表达能力强，且学习和推理算法精细，马尔科夫逻辑网引起了许多研究者的关注，并且已经广泛应用到许多问题上。为了能够有效的使用类别之间的关联信息，本发明使用马尔科夫逻辑网来对中文微博文本建立模型，从而对其进行情绪分类。

[0041] 本发明使用对角牛顿判别式学习方法来学习马尔科夫逻辑表示各个规则的权重，然后使用 marginal 推理方法计算微博文本的情绪类别。

[0042] 以下结合附图描述根据本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法及系统。

[0043] 图 1 是根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法的流程图。图 2 是根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法的实施步骤图。

[0044] 如图 1 和图 2 所示，根据本发明一个实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法，包括如下步骤：

[0045] S101：从网络中获取微博数据。

[0046] 例如：利用网络爬虫充微博页面中抓取微博数据。具体而言，从互联网上获取原创的微博文本（即微博数据），以便后续对微博文本进行情绪分类。在本发明的一个实施例中，微博数据是基于新浪微博开放平台的 API，通过网络爬虫从新浪微博上爬取得到的，并保存到相应的数据库中。所抓取的微博数据为微博文本，如果要对某一事件或某个人的相关微博或评论进行分析，可以使用相应的微博 API 抓取到相关的微博数据。

[0047] S102：根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从所述微博数据中提取相应的特征词，以及根据特征词构建 MLNs 知识库。其中，谓词和规则表示所述微博数据以及微博数据之间的关联关系。

[0048] 具体而言，该步骤主要是根据马尔科夫逻辑网络的基础知识，确定能够表示文本情绪分类的谓词及规则，并根据外部情感情绪词典，从微博文本的数据中提取相应的特征词，构建 MLNs 知识库。文本情绪分类的谓词及规则可以简洁、规整、全面的表示所有的微博文本及其之间的关联关系，外部情感情绪词典中词的全面性、重要性可以构建更为全面的 MLNs 知识库。这样，可以提升分类效果以外，还可以提高分类速度。

[0049] 另外，在实际应用中，还可根据微博文本的特点，灵活地调整马尔科夫逻辑表示，增删外部情感情绪词，即可选用不同的谓词、规则及特征词等。

[0050] S103：根据 MLNs 知识库、谓词和规则训练分类器，对微博数据进行情绪分类并输出中文微博情绪分类结果。

[0051] 具体地说，根据 MLNs 知识库、谓词和规则训练分类器，包括：根据训练数据集训练所述分类器，其中，在训练分类器时，根据证据谓词和查询谓词进行规则参数的学习；根据测试数据集测试分类器的性能，其中，在测试分类器的性能时，采用准确率、召回率和 F 值作为分类器的性能评估指标。也就是说，使用马尔科夫逻辑表示中的规则及根据源数据构建而成的知识库训练分类器。数据分成训练数据集和测试数据集。分类器在训练数据集上训练，在测试数据集上测试分类器性能。训练分类器时是根据证据谓词和查询谓词来学习规则的参数，测试分类器时则是根据证据谓词来预测查询谓词。分类性能的评估指标采用准确率 (Precision)、召回率 (Recall) 和 F 值 (F-measure)。

[0052] 通过本发明实施例的方法对微博数据进行分类的结果如表 2 所示，即通过本发明实施例的方法在微博数据上的分类测试结果，其中，所使用的微博数据是从新浪微博抓取的原创微博文本，共 9960 条。

[0053] 表 2

[0054]

情绪	准确率	召回率	F 值
高兴	93.52%	88.91%	91.15%
悲伤	89.25%	90.22%	89.71%
厌恶	88.44%	74.50%	80.85%
愤怒	60.34%	73.89%	66.34%
恐惧	53.67%	75.46%	62.57%
惊奇	38.30%	70.42%	49.40%

[0055] 根据本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类方法，通过马尔科夫逻辑网络，利用其参数学习和推理体系，对微博文本进行建模，然后将其分成例如高兴、悲伤、厌恶、愤怒、恐惧、惊奇等六类情绪类别，具有分类准确的优点。

[0056] 如图 3 所示，进一步地，本发明的实施例公开了一种基于马尔科夫逻辑网络的中文微博情绪分类系统 300，包括：数据获取模块 310、马尔科夫逻辑表示模块 320 和分类器训练模块 330。

[0057] 其中，数据获取模块 310 用于从网络中获取微博数据。马尔科夫逻辑表示模块 320 用于根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从微博数据中提取相应的特征词，以及根据特征词构建 MLNs 知识库。分类器训练模块 330 用于根据 MLNs 知识库、谓词和规则训练分类器，对微博数据进行情绪分类并输出中文微博情绪分类结果。

[0058] 在本发明的一个实施例中，数据获取模块 310 用于利用网络爬虫充微博页面中抓取微博数据。

[0059] 在本发明的一个实施例中，谓词和规则表示微博数据以及微博数据之间的关联关系。

[0060] 在本发明的一个实施例中，微博数据包括训练数据集和测试数据集。进一步地，分类器训练模块 330 用于 ：根据训练数据集训练分类器，其中，在训练分类器时，根据证据谓词和查询谓词进行规则参数的学习 ；根据测试数据集测试所述分类器的性能，其中，在测试分类器的性能时，采用准确率、召回率和 F 值作为分类器的性能评估指标。

[0061] 根据本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类系统，通过马尔科夫逻辑网络，利用其参数学习和推理体系，对微博文本进行建模，然后将其分成例如高兴、悲伤、厌恶、愤怒、恐惧、惊奇等六类情绪类别，具有分类准确的优点。

[0062] 需要说明的是，本发明实施例的基于马尔科夫逻辑网络的中文微博情绪分类系统的具体实现方式与本发明实施例基于马尔科夫逻辑网络的中文微博情绪分类方法的具体实现方式类似，具体请参见方法部分的描述，为了减少冗余，不做赘述。

[0063] 尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在不脱离本发明的原理和宗旨的情况下在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

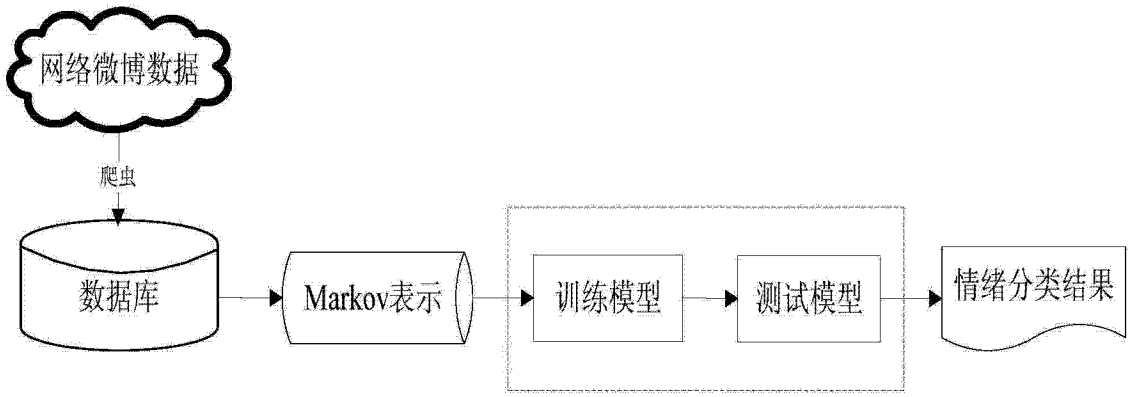

网络微博数据
爬虫
数据库
Markov表示
训练模型
测试模型
情绪分类结果

图 1

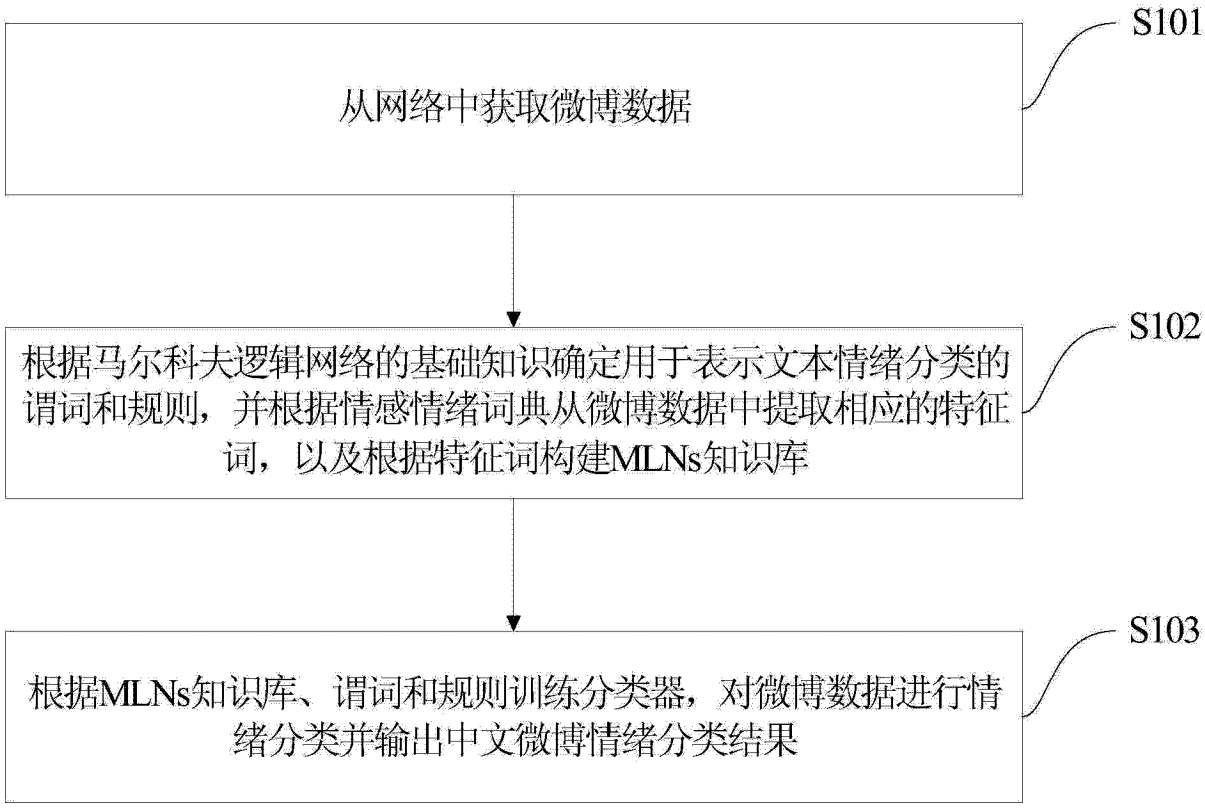

S101
从网络中获取微博数据
S102
根据马尔科夫逻辑网络的基础知识确定用于表示文本情绪分类的谓词和规则，并根据情感情绪词典从微博数据中提取相应的特征词，以及根据特征词构建MLNs知识库
S103
根据MLNs知识库、谓词和规则训练分类器，对微博数据进行情绪分类并输出中文微博情绪分类结果

图 2

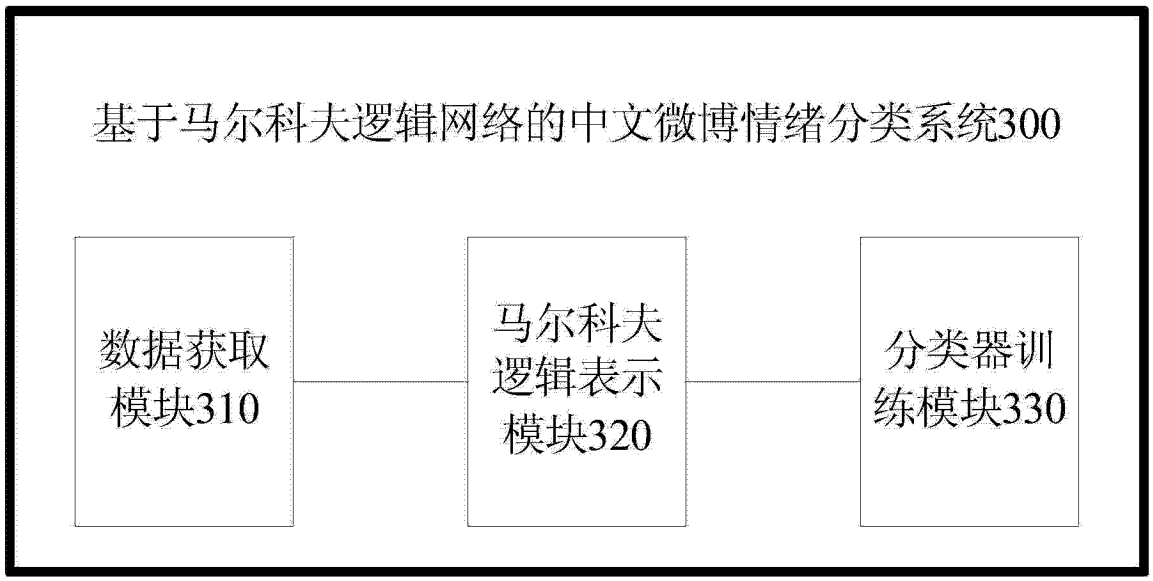

基于马尔科夫逻辑网络的中文微博情绪分类系统300
数据获取模块310
马尔科夫逻辑表示模块320
分类器训练模块330

图 3