(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 106874253 A
(43)申请公布日 2017.06.20

(21)申请号 201510919548.0

(22)申请日 2015.12.11

(71)申请人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市福田区振兴路赛格科技园2栋东403室

(72)发明人 付星辉

(74)专利代理机构 广州三环专利代理有限公司 44202
代理人 郝传鑫 熊永强

(51)Int.Cl.
*G06F 17/27*(2006.01)
*G06F 17/30*(2006.01)

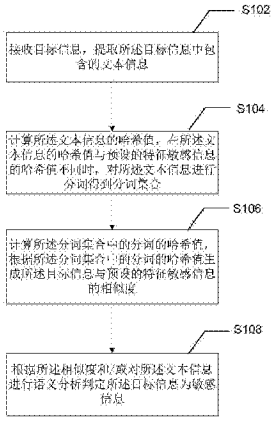
权利要求书2页 说明书9页 附图5页

(54)发明名称

识别敏感信息的方法及装置

(57)摘要

本发明实施例公开了一种识别敏感信息的方法，包括：接收目标信息，提取所述目标信息中包含的文本信息；计算所述文本信息的哈希值，在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合；计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度；根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。本发明还相应地公开了一种识别敏感信息的装置。上述识别敏感信息的方法和装置在对用户发布的内容是否为敏感信息的判定上具有较高的识别准确率。




CN 106874253 A

1.一种识别敏感信息的方法,其特征在于,包括:

接收目标信息,提取所述目标信息中包含的文本信息;

计算所述文本信息的哈希值,在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时,对所述文本信息进行分词得到分词集合;

计算所述分词集合中的分词的哈希值,根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度;

根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

2.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度的步骤包括:

计算所述分词集合中,与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例;

根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

3.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度的步骤包括:

结合simhash算法,根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值;

计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值;

根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

4.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述提取所述目标信息中包含的文本信息的步骤之后还包括:

在所述目标信息中不包含文本信息时,获取发布所述目标信息的用户标识;

获取所述用户标识的行为特征数据,根据所述行为特征数据判定所述目标信息是否为敏感信息。

5.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述计算所述文本信息的哈希值的步骤之后还包括:

在所述文本信息的哈希值与预设的特征敏感信息的哈希值相同时,判定所述目标信息为敏感信息。

6.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息的步骤还包括:

根据预设的机器学习概率模型提取所述文本信息的文本特征;

将所述文本特征作为输入,根据所述预设的机器学习概率模型通过计算所述目标信息的敏感置信度对所述文本信息进行语义分析;

根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

7.根据权利要求6所述的一种识别敏感信息的方法,其特征在于,所述根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息的步骤之后还包括:

若所述目标信息被判定为敏感信息,则将所述目标信息作为特征敏感信息存储。

8.根据权利要求1所述的一种识别敏感信息的方法,其特征在于,所述提取所述目标信息中包含的文本信息的步骤之后还包括:

过滤掉所述文本信息中的符号信息和冗余语义信息。

9.一种识别敏感信息的装置，其特征在于，包括：

文本信息提取模块，用于接收目标信息，提取所述目标信息中包含的文本信息；

全文哈希识别模块，用于计算所述文本信息的哈希值；

分词模块，用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合；

相似度计算模块，用于计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度；

敏感信息判定模块，用于根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

10.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述相似度计算模块还用于计算所述分词集合中，与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例；

根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

11.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述相似度计算模块还用于结合simhash算法，根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值；计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值；根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

12.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述装置还包括行为识别模块，用于在所述目标信息中不包含文本信息时，获取发布所述目标信息的用户标识；获取所述用户标识的行为特征数据，根据所述行为特征数据判定所述目标信息是否为敏感信息。

13.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述全文哈希识别模块还用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值相同时，计算所述文本信息的哈希值，判断所述文本信息的哈希值是否与预设的特征敏感信息的哈希值相同，若是，则判定所述目标信息为敏感信息。

14.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述装置还包括语义识别模块，用于根据预设的机器学习概率模型提取所述文本信息的文本特征；将所述文本特征作为输入，根据所述预设的机器学习概率模型通过计算所述目标信息的敏感置信度对所述文本信息进行语义分析；

所述敏感信息判定模块还用于根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

15.根据权利要求14所述的一种识别敏感信息的装置，其特征在于，所述语义识别模块还用于在所述目标信息被判定为敏感信息时，则将所述目标信息作为特征敏感信息存储。

16.根据权利要求9所述的一种识别敏感信息的装置，其特征在于，所述文本信息提取模块还用于过滤掉所述文本信息中的符号信息和冗余语义信息。

# 识别敏感信息的方法及装置

## 技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种识别敏感信息的方法及装置。

## 背景技术

[0002] 在现有的web2.0的互联网社交应用中，应用的内容不再由服务器发布和推送，而是更多的由用户自行发布和交互。例如，用户可通过手机拍照分享到网络上发送给其他用户，可以编辑论坛主题、博客、论坛发帖、微博等文本内容分享给其他用户。然而，用户分享的内容可能存在违法或者不符合道德规范的风险，例如，粗口、暴力、淫秽、诈骗等内容，因此，需要对用户发布的内容进行敏感信息的识别和拦截。

[0003] 现有的在线拦截敏感信息的方法中，通常采用较单一的文本相似算法策略如全文md5相似来发现拦截敏感信息，虽然这种方法准确率非常高，但是敏感信息的召回率严重依赖于已有的敏感信息特征库的规模，并且敏感信息极容易出现变种，这种相似算法很难有效的发现相似的文本消息，对敏感信息的发现召回率低，且仅通过人工添加敏感信息特征的方法具有一定的时间滞后性，很难解决消息变种问题。

[0004] 因此，传统技术中的在线拦截敏感信息的方法由于人工添加敏感信息特征具有一定的时间滞后性的原因，使得识别敏感信息的准确度不高，对于变种和近似的敏感信息无法准确地识别。

## 发明内容

[0005] 基于此，为传统技术中的在线拦截敏感信息的方法由于人工添加敏感信息特征具有一定的时间滞后性的原因，使得识别敏感信息的准确度不高的技术问题，特提供了一种识别敏感信息的方法。

[0006] 一种识别敏感信息的方法，包括：

[0007] 接收目标信息，提取所述目标信息中包含的文本信息；

[0008] 计算所述文本信息的哈希值，在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合；

[0009] 计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度；

[0010] 根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

[0011] 在其中一个实施例中，所述根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度的步骤包括：

[0012] 计算所述分词集合中，与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例；

[0013] 根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

[0014] 在其中一个实施例中，所述根据所述分词集合中的分词的哈希值生成所述目标信

息与预设的特征敏感信息的相似度的步骤包括：

[0015] 结合simhash算法，根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值；

[0016] 计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值；

[0017] 根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

[0018] 在其中一个实施例中，所述提取所述目标信息中包含的文本信息的步骤之后还包括：

[0019] 在所述目标信息中不包含文本信息时，获取发布所述目标信息的用户标识；

[0020] 获取所述用户标识的行为特征数据，根据所述行为特征数据判定所述目标信息是否为敏感信息。

[0021] 在其中一个实施例中，所述计算所述文本信息的哈希值的步骤之后还包括：

[0022] 在所述文本信息的哈希值与预设的特征敏感信息的哈希值相同时，判定所述目标信息为敏感信息。

[0023] 在其中一个实施例中，所述根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息的步骤还包括：

[0024] 根据预设的机器学习概率模型提取所述文本信息的文本特征；

[0025] 将所述文本特征作为输入，根据所述预设的机器学习概率模型通过计算所述目标信息的敏感置信度对所述文本信息进行语义分析；

[0026] 根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

[0027] 在其中一个实施例中，所述根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息的步骤之后还包括：

[0028] 若所述目标信息被判定为敏感信息，则将所述目标信息作为特征敏感信息存储。

[0029] 在其中一个实施例中，所述提取所述目标信息中包含的文本信息的步骤之后还包括：

[0030] 过滤掉所述文本信息中的符号信息和冗余语义信息。

[0031] 此外，为传统技术中的在线拦截敏感信息的方法由于人工添加敏感信息特征具有一定的时间滞后性的原因，使得识别敏感信息的准确度不高的技术问题，特提供了一种识别敏感信息的装置。

[0032] 一种识别敏感信息的装置，包括：

[0033] 文本信息提取模块，用于接收目标信息，提取所述目标信息中包含的文本信息；

[0034] 全文哈希识别模块，用于计算所述文本信息的哈希值；

[0035] 分词模块，用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合；

[0036] 相似度计算模块，用于计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度；

[0037] 敏感信息判定模块，用于根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

[0038] 在其中一个实施例中，所述相似度计算模块还用于计算所述分词集合中，与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例；

[0039] 根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

[0040] 在其中一个实施例中，所述相似度计算模块还用于结合simhash算法，根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值；计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值；根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

[0041] 在其中一个实施例中，所述装置还包括行为识别模块，用于在所述目标信息中不包含文本信息时，获取发布所述目标信息的用户标识；获取所述用户标识的行为特征数据，根据所述行为特征数据判定所述目标信息是否为敏感信息。

[0042] 在其中一个实施例中，所述全文哈希识别模块还用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值相同时，判定所述目标信息为敏感信息。

[0043] 在其中一个实施例中，所述装置还包括语义识别模块，用于根据预设的机器学习概率模型提取所述文本信息的文本特征；将所述文本特征作为输入，根据所述预设的机器学习概率模型通过计算所述目标信息的敏感置信度对所述文本信息进行语义分析；

[0044] 所述敏感信息判定模块还用于根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

[0045] 在其中一个实施例中，所述语义识别模块还用于在所述目标信息被判定为敏感信息时，则将所述目标信息作为特征敏感信息存储。

[0046] 在其中一个实施例中，所述文本信息提取模块还用于过滤掉所述文本信息中的符号信息和冗余语义信息。

[0047] 实施本发明实施例，将具有如下有益效果：

[0048] 采用了上述识别敏感信息的方法和装置之后，先计算输入的目标信息中的文本信息的哈希值，进行全文哈希比对，使得在目标信息与特征库中的特征敏感信息不完全一致时，可通过对目标信息分词并计算分词的哈希值得到目标信息与特征库中的特征敏感信息的相似度，然后结合对目标信息进行语义分析的分析结果来判定目标信息是否为敏感信息，从而在进行敏感信息的判定时，采用了多种手段，同时结合了全文哈希比对，相似度比对和语义比对的方式，和传统技术相比，即使在目标信息与特征敏感信息不完全相同的情况下，也能够识别出近似的或者变种的敏感信息而不会漏判，从而提高了识别的准确度。

## 附图说明

[0049] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0050] 其中：

[0051] 图1为一个实施例中一种识别敏感信息的方法的流程图；

[0052] 图2为一个实施例中计算目标信息的simhash值的过程示意图；

[0053] 图3为一个实施例中整合了多种识别方式的系统功能图；

[0054] 图4为一个实施例中一种识别敏感信息的装置的示意图；

[0055] 图5为一个实施例中运行前述识别敏感信息的方法的计算机设备的结构示意图。

## 具体实施方式

[0056] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0057] 为传统技术中的在线拦截敏感信息的方法由于人工添加敏感信息特征具有一定的时间滞后性的原因，使得识别敏感信息的准确度不高的技术问题，特提供了一种识别敏感信息的方法，该方法的实现可依赖于计算机程序，该计算机程序可运行于基于冯诺依曼体系的计算机系统之上，该计算机系统可以是社交网络网站或应用、在线游戏网站或应用、在线论坛应用等为用户提供内容发布平台的网站或手机app的服务器。

[0058] 具体的，如图1所示，一种识别敏感信息的方法，包括：

[0059] 步骤S102：接收目标信息，提取所述目标信息中包含的文本信息。

[0060] 如前所述，本申请的应用场景为web2.0应用中，为用户提供内容发布平台的网站或手机app，执行本方法的为用户提供内容发布平台的网站或手机app的服务器。用户通过网页或手机app客户端输入的内容通过终端发送至服务器，再由服务器转发给其他用户，用户输入的内容纪委目标信息。

[0061] 例如，在一个微博应用中，用户通过手机微博客户端拍摄了一张照片，并对该照片添加了文字说明，然后在微博上发布，则该照片和文字说明即为用户发布的内容，服务器接收到的该用户发布的照片和文字说明即为需要判断其是否为包含了暴力、色情、反党反人民、诈骗、传销等不良信息的敏感信息的目标信息。

[0062] 如前所述，用户发布的内容中可不仅包含文本信息，也可以包含图片、音频和视频信息。在本实施例中，可根据目标信息中包含的内容的多媒体类型选择相应的敏感信息识别手段，即若目标信息中包含了文本信息，则对文本信息进行识别，若目标信息中不包含文本信息，例如仅上传了一张图片或一个视频，则根据发布该目标信息的用户的行为特征进行判断。

[0063] 也就是说，在提取所述目标信息中包含的文本信息的步骤之后，服务器可在所述目标信息中不包含文本信息时，获取发布所述目标信息的用户标识；获取所述用户标识的行为特征数据，根据所述行为特征数据判定所述目标信息是否为敏感信息。

[0064] 即可获取该用户标识对应内容发布次数、内容发布频率、被举报次数等行为特征数据计算目标信息为敏感信息的可能性，当该可能性大于阈值时，则判定该目标信息为敏感信息。

[0065] 例如，若某用户在短时间内发布了大量图片，且图片被举报的次数较多，而该用户历史被举报的次数也较多，则服务器可判定该用户发布的内容为敏感信息，从而对其进行屏蔽。

[0066] 而在本实施例中，若可以用户发布的内容中即服务器接收到的目标信息中包含文本内容，则可继续执行后续的步骤S104。

[0067] 步骤S104：计算所述文本信息的哈希值，在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合。

[0068] 在本实施例中，若文本信息的哈希值与预设的特征敏感信息的哈希值完全相同，则可直接判定目标信息为敏感信息，因为哈希值相同则意味着目标信息与特征库中的特征敏感信息完全相同，识别命中。例如，可直接计算文本信息的MD5校验码或SHA1校验码，然后判断该MD5码或SHA1是否与预存的特征敏感信息的MD5码或SHA1相同，若相同，则表示文本信息与特征敏感信息完全相同，从而快速识别出与预存的特征库中的特征敏感信息完全相同的目标信息。而对于文本信息的哈希值与预设的特征敏感信息的哈希值不同的目标信息，才继续进行后续的识别过程(相似度识别和/或语义识别)，从而对于完全匹配特征敏感信息的目标信息进行快速识别，减少无谓的计算，提高执行效率。

[0069] 在本实施例中，在计算得到的文本信息的哈希值与预设的特征敏感信息的哈希值不同时，则对文本信息进行分词。例如，在一个论坛的应用场景中，用户发布的文本信息可能为文本内容较多的传销广告，则可将该传销广告的文本内容进行分词，得到一个一个的单词，该多个单词则为对提取的文本信息进行分词得到的分词集合。在本实施例中，可使用多种开源或非开源的分词工具进行分词，例如StandardAnalyzer、ChineseAnalyzer、CJKAnalyzer等开源分词工具。

[0070] 优选的，在提取所述目标信息中包含的文本信息之后，在分词的过程中，还可进行预处理，过滤掉所述文本信息中的符号信息和冗余语义信息。例如，用户通过微博发布内容是添加有较多的表情符号或者标点符号，由于表情符号和标点符号基本上不涉及敏感信息，可预先将其过滤，从而减少计算量。另外，对于明显笔误的重复出现的分词，可进行过滤，从而可减少计算量。

[0071] 步骤S106：计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度。

[0072] 目标信息与预设的特征敏感信息的相似度即为目标信息和预设的特征敏感信息在相似程度上的量化表示。在本实施例中，根据分词集合中的分词的哈希值有两种方式计算目标信息与预设的特征敏感信息的相似度。

[0073] 第一种，根据匹配的分词的比例生成相似度，具体为：

[0074] 计算所述分词集合中，与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例；根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

[0075] 例如，若目标信息中包含10个分词，其中有8个分词出现在了特征敏感信息中，也就是说这8个分词的哈希值(例如MD5码)，与特征敏感信息的分词中的8个分词的哈希值相同，则分词集合中有80％的分词为与特征敏感信息的分词相同的分词，则可根据该80％生成相似度。

[0076] 第二种，根据simhash算法计算目标信息的文本信息的simhash值，根据该simhash值生成相似度，具体为：

[0077] 结合simhash算法，根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值；计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值；根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

[0078] 例如，若图2所示，图2展示了目标信息的simhash值的计算过程，其中，预设的simhash值为如图2所示的6位。图2中，1至n为分词集合中第1个至第n个分词，每个分词的权

重系数$W_1$至$W_n$即为分词集合中的分词在目标信息中出现的次数，可预先使用特定的哈希算法计算出每个第1个至第n个分词各自的6位的哈希值，然后以此计算目标信息的simhash值的每一位。

[0079] 当计算到目标信息的simhash值的第i位时，则获取到每个分词的哈希值在其第i位上的值。例如，在图2中，

[0080] 第1个分词的哈希值为:100110；

[0081] 第2个分词的哈希值为:110000；

[0082] ……

[0083] 第n个分词的哈希值为:001001；

[0084] 在计算到目标信息的simhash值的第1位时，则获取到：

[0085] 第1个分词的哈希值在第1位上的值为1；

[0086] 第2个分词的哈希值在第1位上的值为1；

[0087] 第n个分词的哈希值在第1位上的值为0；

[0088] 然后根据每个分词的哈希值在其第i位上的值生成该分词的权重系数的符号，然后求和，即可根据该求和得到数值的正负号得到目标信息的simhash值的第i位的数值。

[0089] 如上例中，由于第1个分词的哈希值在第1位上的值为1；第2个分词的哈希值在第1位上的值为1；第n个分词的哈希值在第1位上的值为0；则根据生成的权重系数$W_1$至$W_n$各自的符号得到的求和表达式为：

[0090] $+W_1+W_2\cdots-W_n$

[0091] 若其大于0，则目标信息的simhash值的第1位为1，否则为0。

[0092] 因此计算得到的simhash值也为一个6位的数值，需要说明的是，simhash值的大小可任意设置，但通常可设置为32位或64位大小。优选的，还可使用minhash算法预先对目标信息进行相似筛选，然后再进行simhash，从而可减少计算量。

[0093] 若计算得到的目标信息的simhash值为111001，而1的特征敏感信息中，与该simhash值最接近的特征敏感信息的simhash值为111000，则二者只差为000001，即可根据000001生成相似度。

[0094] 需要说明的时，上述两种计算相似度的方式并不互斥，可在同一个实施例中使用，也就是说，在生成相似度时，可同时计算与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例以及目标信息的simhash值的差值，然后根据二者的结合(例如加权平均后)生成相似度。

[0095] 步骤S108:根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

[0096] 在本实施例中，由于相似度是一个量化值，因此，可判断相似度是否大于阈值，若是则意味着目标信息中大部分分词与特征敏感信息中的分词雷同，因此相似度较高，此时即可判定目标信息为敏感信息。

[0097] 可选的，对于前述计算得到相似度较低的目标信息，由于人为添加用于参考比对的特征敏感信息的实时性不足，可能存在虽然某个目标信息与已存的特征敏感信息的相似度均较低，且分词的差异也较大，但仍然为敏感信息的可能，因此仍然存在漏识别的风险，为此，可继续进行识别，提高识别准确度。

[0098] 具体的，提取所述目标信息中包含的文本信息的步骤之后，服务器还可进行语义识别，即根据预设的机器学习概率模型提取所述文本信息的文本特征；将所述文本特征作为输入，根据所述预设的机器学习概率模型计算所述目标信息的敏感置信度对所述文本信息进行语义分析；根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

[0099] 文本信息的文本特征可包括与预设的特征关键字匹配的分词以及分词出现的顺序等数据。可预先输入敏感信息的样本进行机器学习，从而服务器建立机器学习概率模型，在该机器学习概率模型训练完毕后，则可将提取的文本特征输入到该机器学习概率模型中计算置信度，当置信度大于阈值时，则表示机器识别成功，从而可判定目标信息为敏感信息，否则，判定该目标信息为非敏感信息。

[0100] 在本实施例中，在判定目标信息是否为敏感信息时，可结合相似度和敏感置信度来判定目标信息是否为敏感信息。例如，可先进行相似度分析，若相似度高于第一阈值，则直接判定目标信息为敏感信息，若相似度低于第二阈值，则直接判定目标信息为非敏感信息，若相似度处于第一阈值和第二阈值之间，则对目标信息进行语义识别。也可以先进行语义识别，若敏感置信度高于第三阈值，则直接判定目标信息为敏感信息，若敏感置信度低于第四阈值，则直接判定目标信息为非敏感信息，若相似度处于第三阈值和第四阈值之间，则对目标信息进行相似度识别。(由于语义识别的机器学习计算量较大，则优选的可先进行相似度识别再进行语义识别)。在另一个实施例中，还可综合相似度与敏感置信度，在二者均满足预设的条件时，识别目标信息为敏感信息。也就是说，在本方法的多个应用场景中，设计人员可根据对敏感信息审查的严格程度自行设置相似度和敏感置信度所需要符合的条件参数，从而可采用不同的松紧度策略对敏感信息进行识别。

[0101] 优选的，根据所述敏感置信度判定所述目标信息是否为敏感信息的步骤之后还包括：

[0102] 若所述目标信息被判定为敏感信息，则将所述目标信息作为特征敏感信息存储。

[0103] 也就是说，若通过机器学习概率模型成功识别出了敏感信息，则将该敏感信息作为比对的参考样本添加到预设的特征库中作为特征敏感信息存储，后续若再有用户输入该内容，则预先通过全文MD5比对则可快速地识别出敏感信息，而不用后续的繁琐的分词比对识别和机器学习识别的过程，从而减少了计算量，提高了执行效率。

[0104] 需要说明的是，如图3所示，在一个整合的敏感信息识别服务器上，上述敏感信息识别的过程包括全文MD5比对的精确识别方式、分词后MD5比对或simhash值比对的相似度识别方式、基于机器学习的语义识别方式，根据发布目标信息的用户标识的行为特征数据进行识别的行为识别方式可应用于同一个系统中。在该系统中，对用户发布的内容先进行预处理，去除语义上冗余重复的分词，去掉标点符号和表情符号，然后根据内容是否包含文本信息选择识别方式，若不包含文本信息则选择行为识别。若包含文本信息，则可先后进行精确识别、相似识别和语义识别三个阶段，最终根据精确识别、相似识别和语义识别的结果判定目标信息是否为敏感信息。此种整合了多种识别方式的方法或系统由于从多个维度进行识别，因此有着较高的识别准确度。

[0105] 为传统技术中的在线拦截敏感信息的方法由于人工添加敏感信息特征具有一定的时间滞后性的原因，使得识别敏感信息的准确度不高的技术问题，特提供了一种识别敏感信息的装置，如图4所示，该装置包括文本信息提取模块102、全文哈希识别模块104、分词

模块106、相似度计算模块108以及敏感信息判定模块110，其中：

[0106] 文本信息提取模块102，用于接收目标信息，提取所述目标信息中包含的文本信息。

[0107] 全文哈希识别模块104，用于计算所述文本信息的哈希值。

[0108] 分词模块106，用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合。

[0109] 相似度计算模块108，用于计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度。

[0110] 敏感信息判定模块110，用于根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息。

[0111] 在一个实施例中，相似度计算模块106还用于计算所述分词集合中，与预设的特征敏感信息的分词的哈希值匹配的分词在所述分词集合中所占的比例；

[0112] 根据所述比例生成所述目标信息与预设的特征敏感信息的相似度。

[0113] 在一个实施例中，相似度计算模块106还用于结合simhash算法，根据所述分词集合中的分词的哈希值生成所述目标信息的第一simhash值；计算所述第一simhash值与所述预设的特征敏感信息的第二simhash值的差值；根据所述差值生成所述目标信息与预设的特征敏感信息的相似度。

[0114] 在一个实施例中，如图4所示，该装置还包括行为识别模块112，用于在所述目标信息中不包含文本信息时，获取发布所述目标信息的用户标识；获取所述用户标识的行为特征数据，根据所述行为特征数据判定所述目标信息是否为敏感信息。

[0115] 在一个实施例中，如图4所示，全文哈希识别模块104还用于在所述文本信息的哈希值与预设的特征敏感信息的哈希值相同时，判定所述目标信息为敏感信息。

[0116] 在一个实施例中，如图4所示，该装置还包括语义识别模块114，用于根据预设的机器学习概率模型提取所述文本信息的文本特征；将所述文本特征作为输入，根据所述预设的机器学习概率模型计算所述目标信息的敏感置信度对所述文本信息进行语义分析；所述敏感信息判定模块还用于根据所述相似度和/或敏感置信度判定所述目标信息是否为敏感信息。

[0117] 在一个实施例中，语义识别模块114还用于在所述目标信息被判定为敏感信息时，则将所述目标信息作为特征敏感信息存储。

[0118] 在一个实施例中，文本信息提取模块102还用于过滤掉所述文本信息中的符号信息和冗余语义信息。

[0119] 实施本发明实施例，将具有如下有益效果：

[0120] 采用了上述识别敏感信息的方法和装置之后，先计算输入的目标信息中的文本信息的哈希值，进行全文哈希比对，使得在目标信息与特征库中的特征敏感信息不完全一致时，可通过对目标信息分词并计算分词的哈希值得到目标信息与特征库中的特征敏感信息的相似度，然后结合对目标信息进行语义分析的分析结果来判定目标信息是否为敏感信息，从而在进行敏感信息的判定时，采用了多种手段，同时结合了全文哈希比对，相似度比对和语义比对的方式，和传统技术相比，即使在目标信息与特征敏感信息不完全相同的情况下，也能够识别出近似的或者变种的敏感信息而不会漏判，从而提高了识别的准确度。

[0121] 在一个实施例中，如图5所示，图5展示了一种运行上述识别敏感信息的方法的基于冯诺依曼体系的计算机系统的终端10。该计算机系统可以是智能手机、平板电脑、掌上电脑，笔记本电脑或个人电脑等终端设备。具体的，可包括通过系统总线连接的外部输入接口1001、处理器1002、存储器1003和输出接口1004。其中，外部输入接口1001可选的可至少包括网络接口10012。存储器1003可包括外存储器10032(例如硬盘、光盘或软盘等)和内存储器10034。输出接口1004可至少包括显示屏10042等设备。

[0122] 在本实施例中，本方法的运行基于计算机程序，该计算机程序的程序文件存储于前述基于冯诺依曼体系的计算机系统10的外存储器10032中，在运行时被加载到内存储器10034中，然后被编译为机器码之后传递至处理器1002中执行，从而使得基于冯诺依曼体系的计算机系统10中形成逻辑上的文本信息提取模块102、全文哈希识别模块104、分词模块106、相似度计算模块108以及敏感信息判定模块110。且在上述识别敏感信息的方法执行过程中，输入的参数均通过外部输入接口1001接收，并传递至存储器1003中缓存，然后输入到处理器1002中进行处理，处理的结果数据或缓存于存储器1003中进行后续地处理，或被传递至输出接口1004进行输出。

[0123] 以上所揭露的仅为本发明较佳实施例而已，当然不能以此来限定本发明之权利范围，因此依本发明权利要求所作的等同变化，仍属本发明所涵盖的范围。

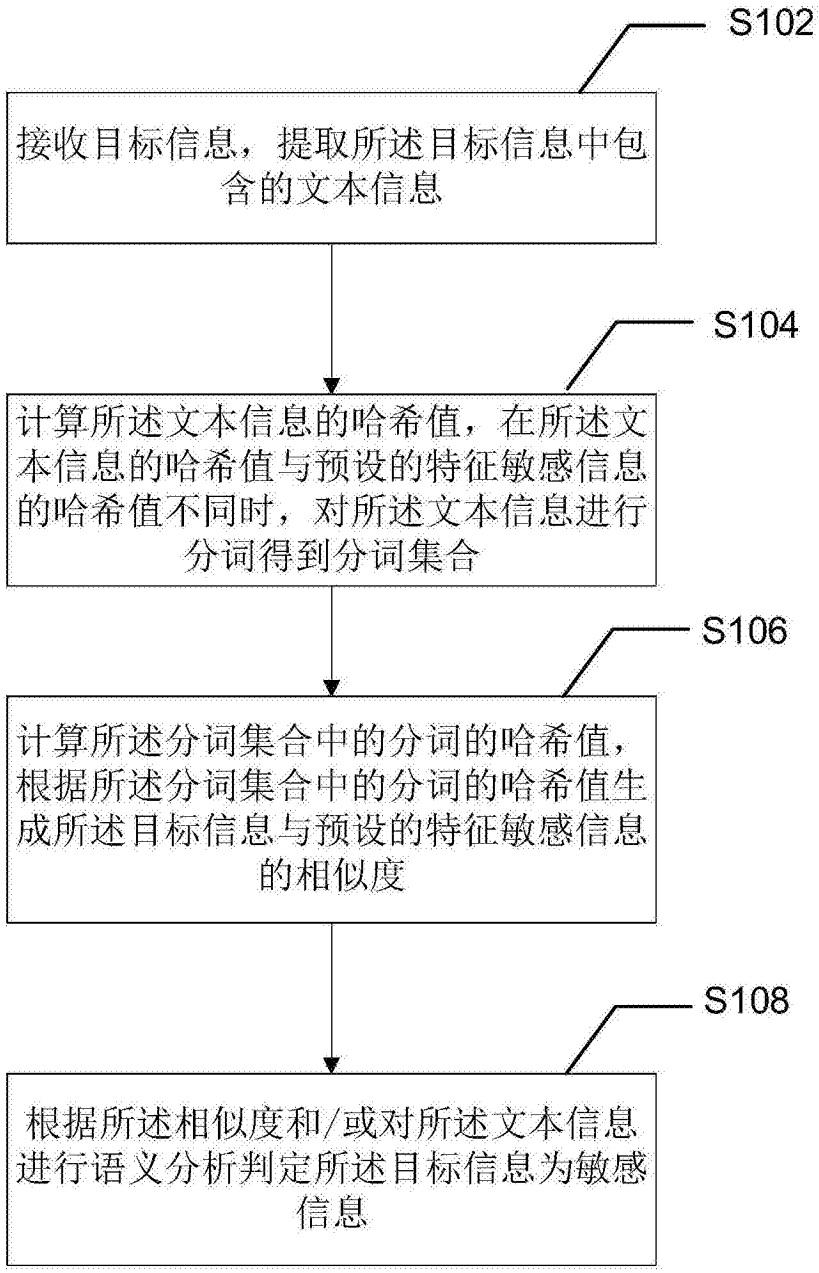

S102
接收目标信息，提取所述目标信息中包含的文本信息
S104
计算所述文本信息的哈希值，在所述文本信息的哈希值与预设的特征敏感信息的哈希值不同时，对所述文本信息进行分词得到分词集合
S106
计算所述分词集合中的分词的哈希值，根据所述分词集合中的分词的哈希值生成所述目标信息与预设的特征敏感信息的相似度
S108
根据所述相似度和/或对所述文本信息进行语义分析判定所述目标信息为敏感信息

图1

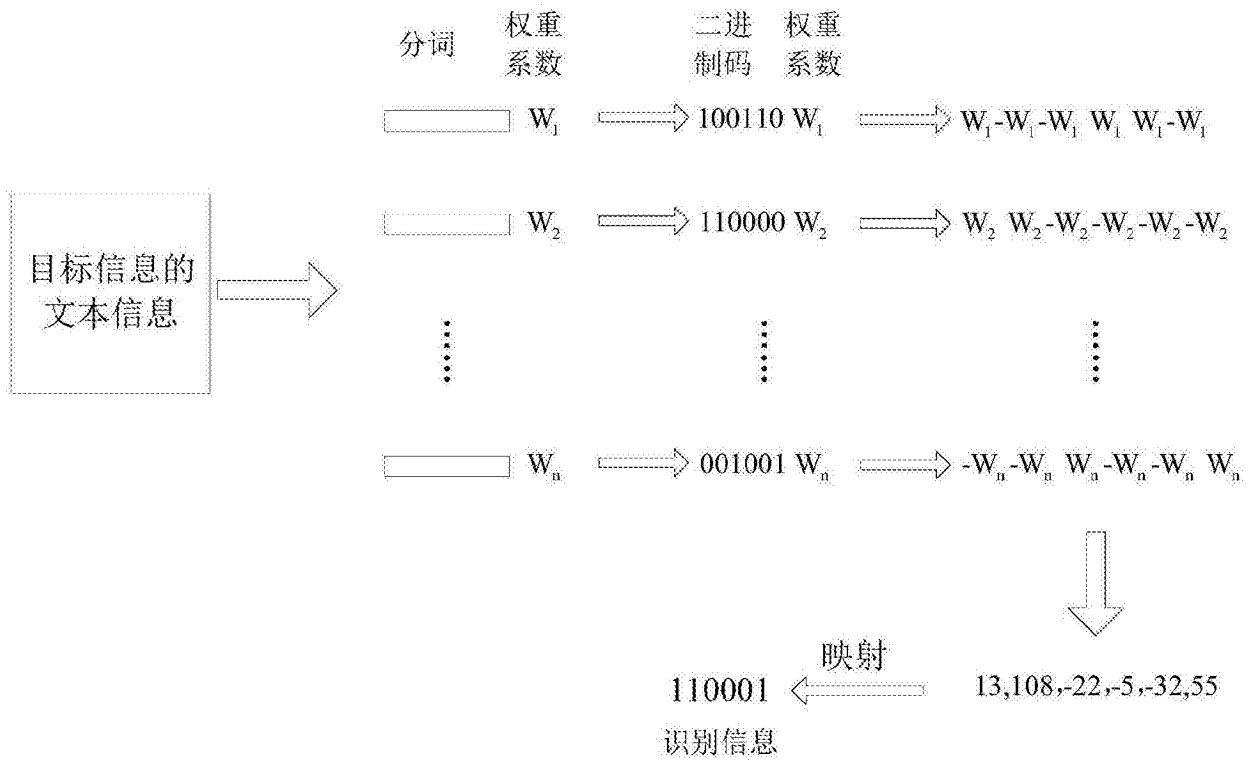

分词
权重系数
二进制码
权重系数
目标信息的文本信息
W1
100110 W1
W1-W1-W1 W1 W1-W1
W2
110000 W2
W2 W2-W2-W2-W2-W2
Wn
001001 Wn
-Wn-Wn Wn-Wn-Wn Wn
映射
110001
识别信息
13,108,-22,-5,-32,55

图2

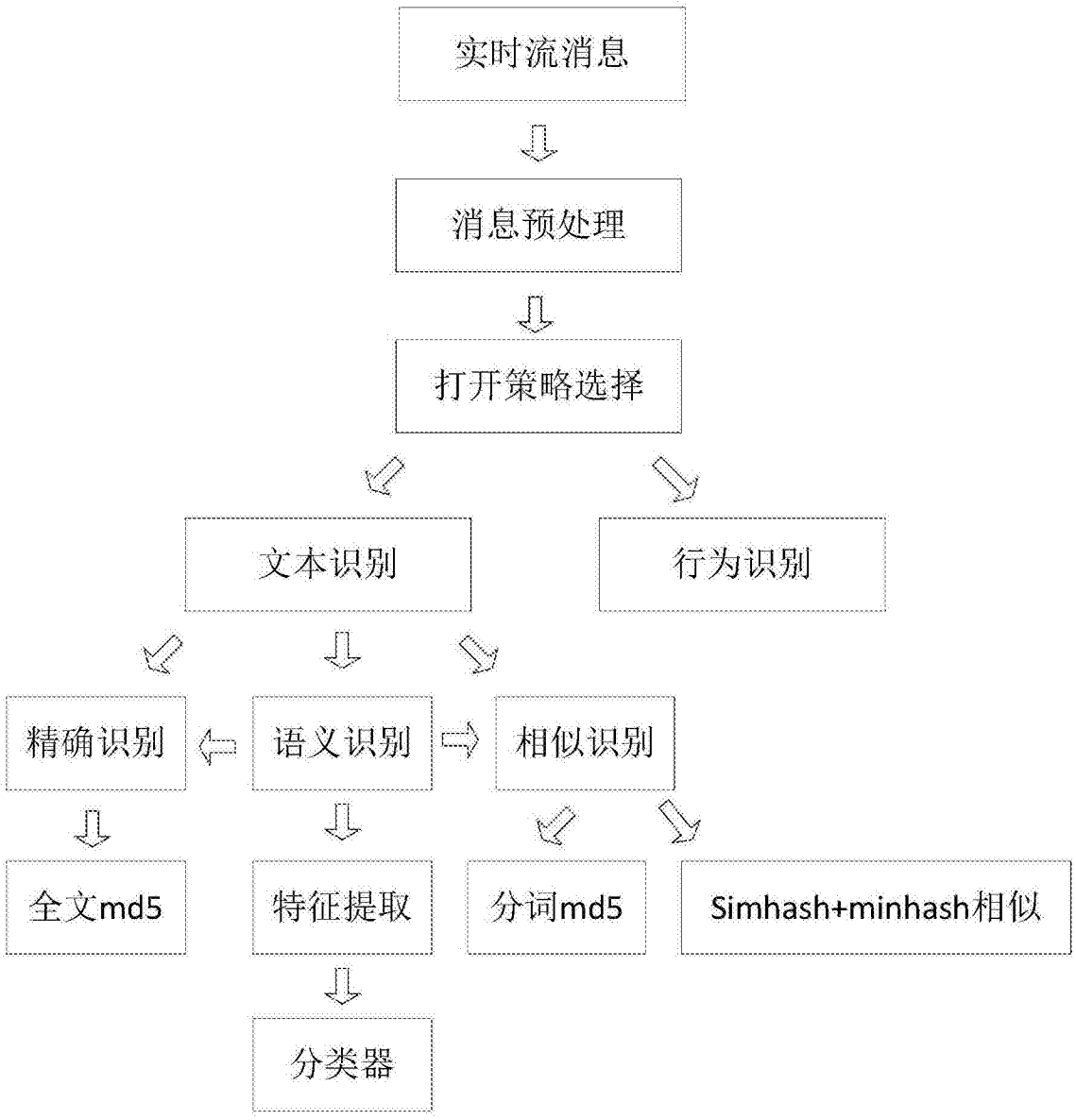

实时流消息
消息预处理
打开策略选择
文本识别
行为识别
精确识别
语义识别
相似识别
全文md5
特征提取
分词md5
Simhash+minhash相似
分类器

图3

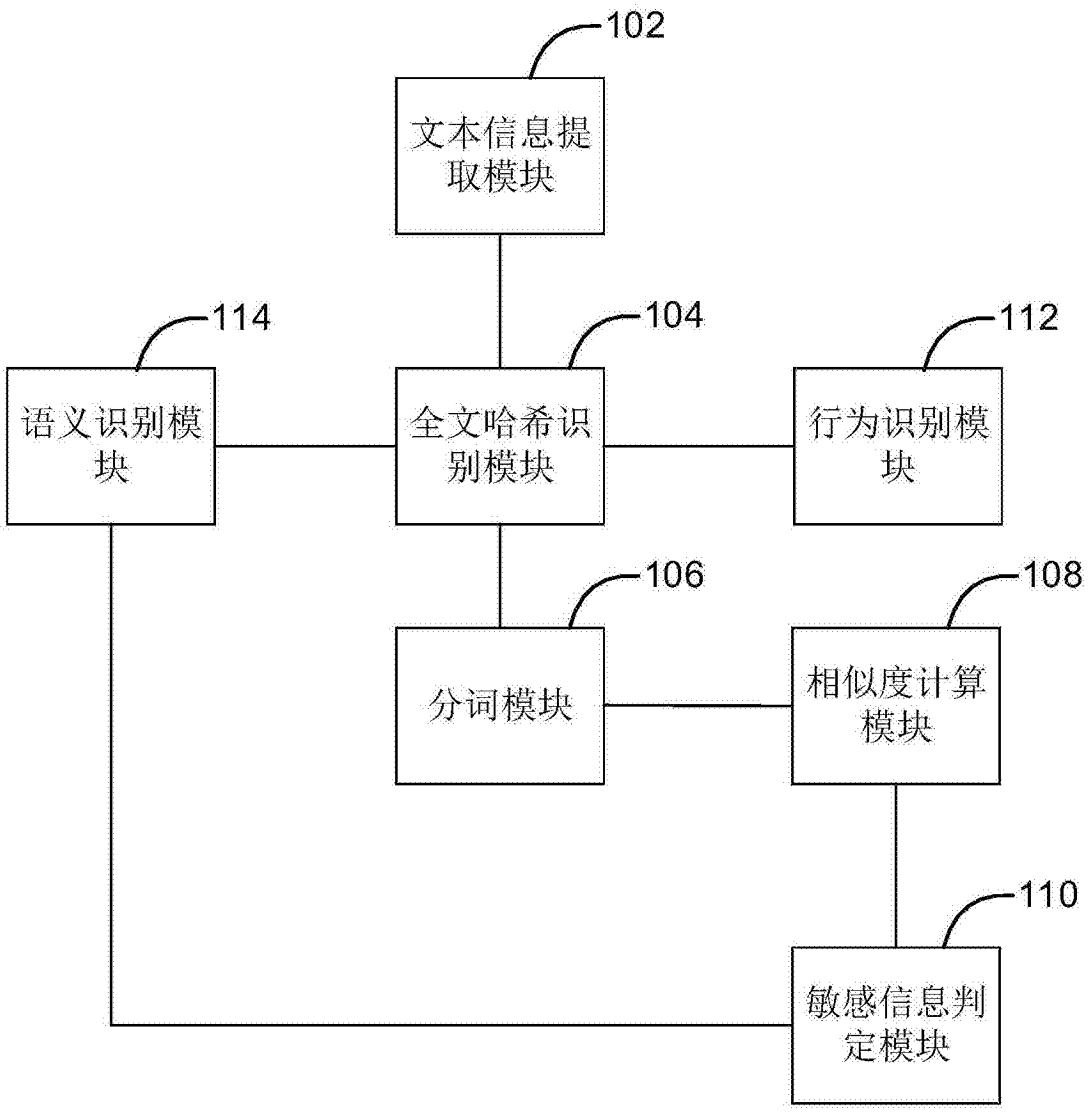

102
文本信息提取模块
114
语义识别模块
104
全文哈希识别模块
112
行为识别模块
106
分词模块
108
相似度计算模块
110
敏感信息判定模块

图4

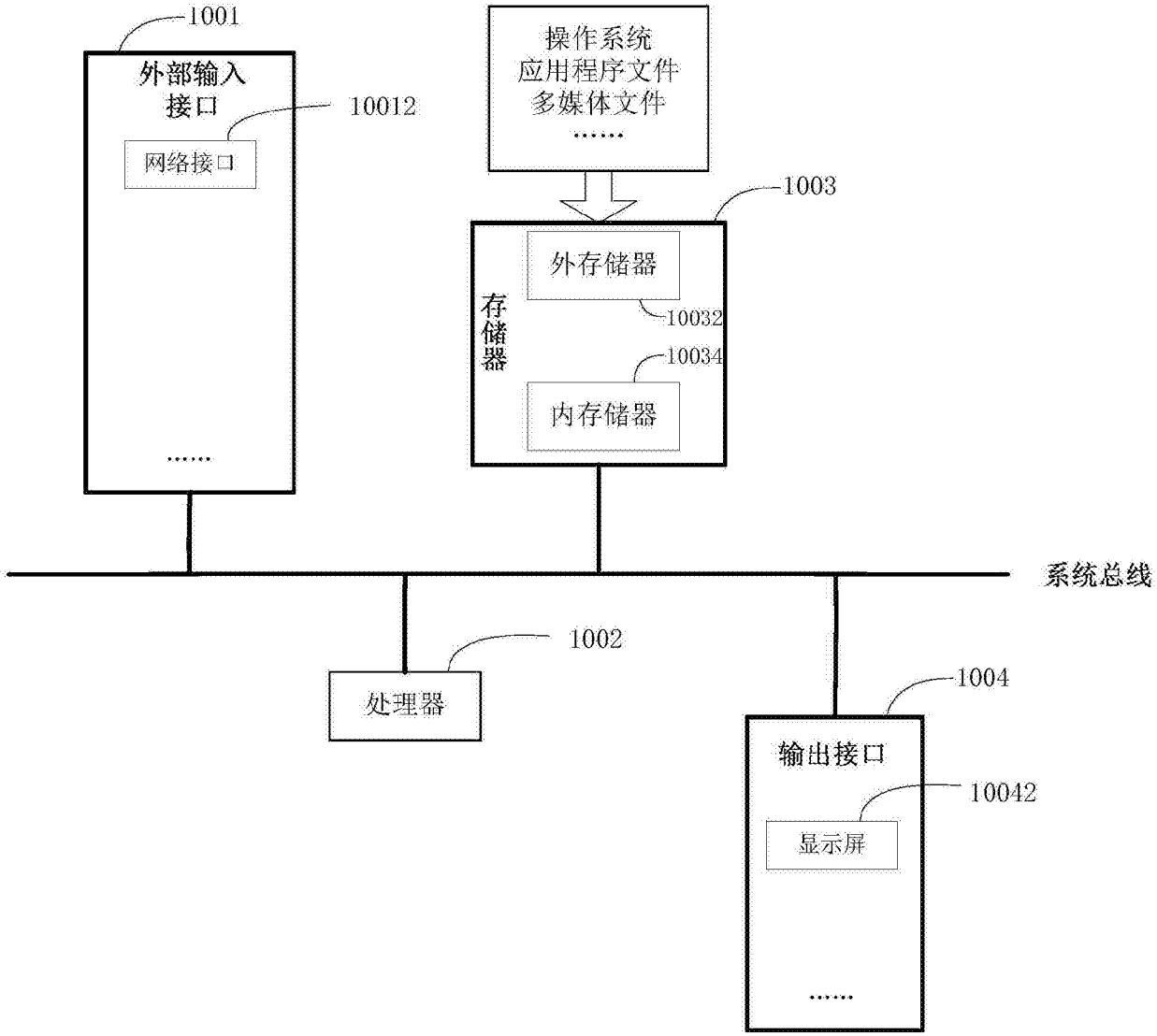

1001
外部输入
接口
10012
网络接口
.....
操作系统
应用程序文件
多媒体文件
.....
1003
外存储器
10032
存
储
器
10034
内存储器
系统总线
1002
处理器
1004
输出接口
10042
显示屏
.....

图5