



(12) 发明专利

(10) 授权公告号 CN 102567304 B

(45) 授权公告日 2014. 02. 26

(21) 申请号 201010621142. 1

(22) 申请日 2010. 12. 24

(73) 专利权人 北大方正集团有限公司

地址 100871 北京市海淀区成府路 298 号方正大厦 5 层

专利权人 北京大学

北京北大方正电子有限公司

(72) 发明人 郑妍 于晓明 杨建武

(74) 专利代理机构 北京中博世达专利商标代理有限公司 11274

代理人 申健

(51) Int. Cl.

G06F 17/27(2006. 01)

(56) 对比文件

US 5987457 A, 1996. 11. 16, 全文.

US 2004/0167964 A1, 2004. 08. 26, 全文.

CN 101477544 A, 2009. 07. 08, 全文.

CN 101877704 A, 2010. 11. 03, 全文.

CN 101894102 A, 2010. 11. 24, 全文.

CN 101908055 A, 2010. 12. 08, 全文.

CN 101794303 A, 2010. 08. 04, 全文.

CN 101639824 A, 2010. 02. 03, 全文.

马建国 等. 信息过滤技术及 Visual J++ 实现. 《系统工程与电子技术》. 2004, 第 26 卷 (第 3 期), 第 382-385 页.

审查员 田涛

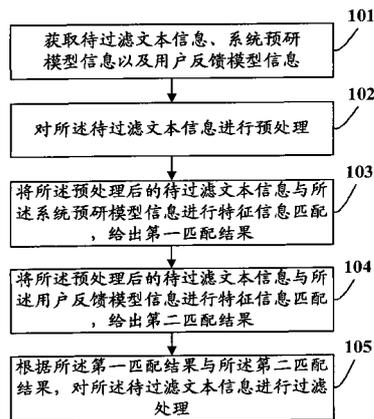
权利要求书3页 说明书8页 附图3页

(54) 发明名称

一种网络不良信息的过滤方法及装置

(57) 摘要

本发明公开了一种网络不良信息的过滤方法及装置,涉及计算机信息处理及信息过滤技术领域。其中,本发明实施例提供的一种网络不良信息的过滤方法,包括:获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息;对所述待过滤文本信息进行预处理;将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果;将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;根据所述第一匹配结果与第二匹配结果,对所述待过滤文本信息进行过滤处理。采用本发明实施例能够实现提高不良信息自动过滤性能,且可以实现系统信息自动更新。



1. 一种网络不良信息的过滤方法,其特征在于,包括:

获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息,所述系统预研模型信息包括:规则索引库和系统预研模型特征项信息;所述用户反馈模型信息包括:规则索引库和用户反馈模型特征项信息;

对所述待过滤文本信息进行预处理;

将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果;

将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;

根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本信息进行过滤处理。

2. 根据权利要求1所述的网络不良信息的过滤方法,其特征在于,该方法还包括:

获取所述系统预研模型信息的语料与所述用户反馈模型信息的语料。

3. 根据权利要求2所述的网络不良信息的过滤方法,其特征在于,所述用户反馈模型信息的语料包括:用户反馈语料和/或被过滤语料。

4. 根据权利要求3所述的网络不良信息的过滤方法,其特征在于,该方法还包括:

获取所述用户反馈模型信息的语料数量以及其对应的阈值;

根据所述用户反馈模型信息的语料数量以及其对应的阈值,对所述用户反馈模型信息进行更新。

5. 根据权利要求2或3或4所述的网络不良信息的过滤方法,其特征在于,所述对所述待过滤文本信息进行预处理的步骤,包括:

对所述待过滤文本信息进行切分处理;

统计所述切分处理后的候选特征项数量。

6. 根据权利要求5所述的网络不良信息的过滤方法,其特征在于,所述将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果步骤,包括:

获取所述预处理后的待过滤文本信息以及所述系统预研模型信息;

将所述预处理后的待过滤文本信息与所述系统预研模型信息进行匹配,获取特征项;

统计所述特征项的语料信息得分;

根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;

根据判断结果,给出所述第一匹配结果。

7. 根据权利要求6所述的网络不良信息的过滤方法,其特征在于,所述将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果步骤,包括:

获取所述预处理后的待过滤文本信息以及所述用户反馈模型信息;

将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行匹配,获取特征项;

统计所述特征项的语料信息得分;

根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;

根据判断结果,给出所述第二匹配结果。

8. 根据权利要求1所述的网络不良信息的过滤方法,其特征在于,所述系统预研模型

信息的规则索引库包括：系统预置规则；所述用户反馈模型信息的规则索引库包括：用户配置规则。

9. 一种网络不良信息的过滤装置，其特征在于，包括：

信息获取单元，用于获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息，所述系统预研模型信息包括：规则索引库和系统预研模型特征项信息；所述用户反馈模型信息包括：规则索引库和用户反馈模型特征项信息；

预处理单元，用于对所述待过滤文本信息进行预处理；

第一匹配单元，用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配，给出第一匹配结果；

第二匹配单元，用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配，给出第二匹配结果；

过滤单元，用于根据所述第一匹配结果与所述第二匹配结果，对所述待过滤文本信息进行过滤处理。

10. 根据权利要求 9 所述的网络不良信息的过滤装置，其特征在于，所述信息获取单元，还用于获取所述用户反馈模型信息的语料。

11. 根据权利要求 10 所述的网络不良信息的过滤装置，其特征在于，所述用户反馈模型信息的语料包括：用户反馈语料和 / 或被过滤语料。

12. 根据权利要求 11 所述的网络不良信息的过滤装置，其特征在于，该方装置还包括：

阈值获取单元，用于获取所述用户反馈模型信息的语料数量以及其对应的阈值；

更新单元，用于根据所述用户反馈模型信息的语料数量以及其对应的阈值，对所述用户反馈模型信息进行更新。

13. 根据权利要求 10 或 11 或 12 所述的网络不良信息的过滤装置，其特征在于，所述预处理单元，包括：

切分子单元，用于对所述待过滤文本信息进行切分处理；

统计子单元，用于统计所述切分处理后的候选特征项数量。

14. 根据权利要求 13 所述的网络不良信息的过滤装置，其特征在于，所述第一匹配单元，包括：

信息获取子单元，用于获取所述预处理后的待过滤文本信息以及所述系统预研模型信息；

匹配子单元，用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行匹配，获取特征项；

统计子单元，用于统计所述特征项的语料信息得分；

判断子单元，用于根据所述语料信息得分，判断所述特征项所对应的待过滤文本信息是否为不良信息；

结果输出子单元，用于根据判断结果，给出所述第一匹配结果。

15. 根据权利要求 14 所述的网络不良信息的过滤装置，其特征在于，所述第二匹配单元，包括：

信息获取子单元，用于获取所述预处理后的待过滤文本信息以及所述用户反馈模型信息；

匹配子单元,用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行匹配,获取特征项;

统计子单元,用于统计所述特征项的语料信息得分;

判断子单元,用于根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;

结果输出子单元,用于根据判断结果,给出所述第二匹配结果。

一种网络不良信息的过滤方法及装置

技术领域

[0001] 本发明涉及计算机信息处理及信息过滤技术领域,尤其涉及一种基于统计与规则的网络不良信息的过滤方法及装置。

背景技术

[0002] 随着互联网的迅速发展,信息传播速度也随之加快。由于互联网上的内容良莠不齐,例如:广告、色情、暴力以及反动为主的不良信息都难以杜绝,并渐渐以更为隐蔽的方式扩散,因此,抑制不良信息的扩散以及净化互联网络空间就显得十分重要。对于互联网中海量的数据信息,如果采用人工的方法去过滤互联网上的不良信息,则需要耗费巨大的人力物力。因此,基于互联网内容的不良信息的自动过滤技术成为近年来研究的热点。

[0003] 目前,基于互联网内容的不良信息自动过滤技术通常采用如下两种方式:

[0004] (1) 基于关键字匹配的过滤方法;该方法在判定过程中,采取精确匹配的策略,过滤掉出现关键字的文本。采用该方法过滤互联网内容的不良信息速度快,简单易操作。

[0005] (2) 基于统计的文本分类模型的过滤方法;该方法中基于统计的不良文本过滤模型本质上是一个两类的文本分类问题,文本分类是自然语言处理领域的研究重点方向,有大量经典模型可供参考。基于统计的文本分类模型从理论角度来看应该是效果不错的方法,但在实际应用中性能却不理想,误判情况十分突出,主要原因分析如下:

[0006] (1) 正向与负向语料不均衡。其中,正向语料只包含了少量类别,例如:广告、色情、暴力、反动以及用户所关心的不良信息为主。负向语料则包含了大量类别,例如:按照文本内容可划分为:经济、体育、政治、医药、艺术、历史、政治、文化、环境、交通、计算机、教育、军事等等。

[0007] (2) 不良信息的内容表现具有很大的多变性和隐蔽性。发布者经常有意避开常用词,取而代之,如:同音字,拆分字,非汉字噪音,缩略现象,新词等。

[0008] (3) 用户词典只提供关键词精确匹配方式,造成判定方法的机械与不灵活。且单一关键词的语义倾向性不具有代表性,误判率高。比如,当“免费”和“发票”同时出现在上下文环境中要比单一的“发票”更具有说服力。

[0009] (4) 一些传统的中文信息处理做法并不适用于基于文本分类的不良信息过滤。如使用一定规模的禁用词;如特征项只包括双字以上的词汇等。

[0010] (5) 缺少统一的模型,对包括广告、色情、暴力、反动等不良信息进行综合过滤。

[0011] 在实现上述基于互联网内容的不良信息自动过滤技术的过程中,发明人发现现有技术中,不良信息自动过滤性能无法满足当前互联网的过滤需求,且无法实现自动更新。

发明内容

[0012] 本发明实施例提供一种网络不良信息的过滤方法及装置,为达到上述目的,本发明的实施例采用如下技术方案:

[0013] 一种网络不良信息的过滤方法,包括:

- [0014] 获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息；
- [0015] 对所述待过滤文本信息进行预处理；
- [0016] 将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配，给出第一匹配结果；
- [0017] 将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配，给出第二匹配结果；
- [0018] 根据所述第一匹配结果与所述第二匹配结果，对所述待过滤文本信息进行过滤处理。
- [0019] 一种网络不良信息的过滤装置，包括：
- [0020] 信息获取单元，用于获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息；
- [0021] 预处理单元，用于对所述待过滤文本信息进行预处理；
- [0022] 第一匹配单元，用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配，给出第一匹配结果；
- [0023] 第二匹配单元，用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配，给出第二匹配结果；
- [0024] 过滤单元，用于根据所述第一匹配结果与所述第二匹配结果，对所述待过滤文本信息进行过滤处理。
- [0025] 本发明实施例提供的网络不良信息的过滤方法以及装置，通过获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息；对所述待过滤文本信息进行预处理；将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配，给出第一匹配结果；将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配，给出第二匹配结果；根据所述第一匹配结果与所述第二匹配结果，对所述待过滤文本信息进行过滤处理。由于本发明中采用了两次匹配进行系统过滤，所以系统自动过滤不良信息的准确性较高，从而提高了系统的性能；还由于本发明实施例采用了用户反馈模型信息进行不良信息的过滤，使得用户反馈信息能够及时的应用于系统自动过滤不良信息的过程中，从而实现了系统模型信息自动更新的功能。

附图说明

- [0026] 图 1 为本发明实施例提供的一种网络不良信息的过滤方法流程图；
- [0027] 图 2 为本发明实施例提供的另一种网络不良信息的过滤方法流程图；
- [0028] 图 3 为本发明实施例提供的一种网络不良信息的过滤装置结构示意图；
- [0029] 图 4 为本发明实施例提供的另一种网络不良信息的过滤装置结构示意图。

具体实施方式

[0030] 下面结合附图对本发明实施例提供的一种网络不良信息的过滤方法以及装置进行详细描述。

[0031] 如图 1 所述，为本发明实施例提供的一种网络不良信息的过滤方法；该方法包括：

[0032] 101：获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息；

[0033] 102 :对所述待过滤文本信息进行预处理 ;

[0034] 103 :将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果 ;

[0035] 104 :将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果 ;

[0036] 105 :根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本信息进行过滤处理。

[0037] 如图 2 所述,为本发明实施例提供的另一种网络不良信息的过滤方法,该方法包括 :

[0038] 201 :获取所述系统预研模型信息的语料与所述用户反馈模型信息的语料。其中,所述用户反馈模型信息的语料可以包括 :用户反馈语料和 / 或被过滤语料。通常所述系统预研模型与所述用户反馈模型的训练语料的选择分为正向语料与负向语料 ;例如 :正向语料的准备的不良信息内容文本的收集可以主要包括 :广告、色情、暴力、反动等内容文本,共 10000 篇 ;负向语料的准备的非不良信息内容文本的收集主要包含任务主要的文本类别,如经济、政治、体育、文化、医药、交通、环境、军事、文艺、历史、计算机、教育、法律、房产、科技、汽车、人才、娱乐等,共 30000 篇。

[0039] 需要说明的是,所述训练语料的收集过程中,经常出现正负语料不均衡 ;一个类别的语料范围很广,另一个类别语料范围则相对较少。本发明中的解决方案是允许这种不均衡的语料分布,对于语料范围很大的类别的准备策略是求全不求量。

[0040] 202 :获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息 ;

[0041] 203 :对所述待过滤文本信息进行预处理 ;

[0042] 该步骤具体包括 :对所述待过滤文本信息进行切分处理 ;例如 :根据标点和常见词对语料进行断句,常见词是指常用且对判定无意义的词汇,如“的”、“了”等,但“您”较常见于正向语料,“我”较常见于负向语料,不适合作为常用词。

[0043] 需要注意的是,自然语言处理中常用的禁用词表不合作为常用词表。通常可采用方正智思分词 4.0 对语料进行分词及词性标注工作。所述切分处理后的切分单元是后续工作最小的处理单元。

[0044] 统计所述切分处理后的候选特征项数量。例如 :对所述切分处理后的切分单元统计其中非汉字部分数量 ;如 :所述切分单元总数为 N_1 ,非汉字部分为 N_2 ,若 N_2/N_1 大于阈值,则判定此候选特征项所对应的待滤文本信息为不良信息。依据是此信息中含有大量噪音字符,可能是广告等垃圾文本 ;或者,统计所述切分单元中的网址、电话、邮箱、QQ 等联系方式出现数量 $\text{num}(\text{ad})$,此类信息常用于广告中,并赋予默认权重 SCORE_{ad} 。

[0045] 204 :将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果 ;该步骤具体可以包括 :

[0046] 2041 :获取所述预处理后的待过滤文本信息以及所述系统预研模型信息 ;所述系统预研模型信息包括 :规则索引库和所述系统预研模型特征项信息 ;其中,所述规则索引库中的用户规则索引库和用户关键词索引库的生成过程如下 :

[0047] 步骤 S1 :关键词解析 ;所述关键词解析方法为 :首先,对常用字的汉语拼音建索引,依据关键词中每个字的汉语拼音索引生成整体关键词的索引 ;然后,再对关键词中的每

个字进行结构上的拆分,依据拆分结果递归重组关键词;最后,将关键词索引及拆分集合形成键值对,保存所有解析结果生成用户关键词索引库。如“法轮功”关键词解析后,会生成一个索引值,且有多种拆分结果,具体可以包括,“三去车仑工力”,“法车仑功”等等。

[0048] 步骤 S2:语法解析;计算机将规则语法解析为能够处理的形式。所述规则语法包括:AND、OR、NEAR、NOT。如“A ANDB”,其中 A 与 B 都是待解析的关键词,AND 语法表示在上下文环境中,当 A 与 B 同时出现时,该条规则匹配成功。对关键词及规则语法形成键值对,保存所有解析结果生成用户规则索引库。

[0049] 需要注意的是,以上所述索引库规则可以是用户配置的规则,还可以系统预置规则;以上所述步骤是对用户配置规则进行解析生成相应的索引库过程,该索引库可以优化以下匹配过程。

[0050] 2042:将所述预处理后的待过滤文本信息与所述系统预研模型信息进行匹配,获取特征项;其中,所述系统预研模型信息包括:规则索引库和所述系统预研模型特征项信息;该步骤获取系统预研模型特征项信息的过程具体可以为:

[0051] 步骤 S1,将所述切分单元组成词串作为候选特征项;例如:

[0052] (1) 对连续的切分单元组合成词串。对于每句中的切分单元,从第 1 个切分单元开始,组合窗口最大为 N,进行组合。如有序切分单元“ABCD”,最大窗口为 3,则生成词串的组合共有 9 种:ABC、BCD、AB、BC、CD、A、B、C、D。

[0053] (2) 对非连续的切分单元组合成词串。对(1)中的生成的词串计算汉语拼音索引,依据所述 2041 中的步骤 S1 生成的用户关键词索引库中进行匹配。若有匹配成功的集合,统计匹配成功数量 $\text{num}(\text{user})$;然后,再依据所述 2041 中的步骤 S2 生成的用户规则索引库中进行匹配,若匹配成功,对于非连续的切分单元生成一个词串。如(1)中 9 个词串,若在用户关键词索引库中匹配成功两个词串 A、D。在用户规则索引库中有规则“A NEAR2 D”,则生成新的特征项 AD。这里的 2 代表 A 与 D 的距离不超过 2。累加统计匹配成功数量 $\text{num}(\text{user})$,赋予默认权重 $\text{SCOFF}_{\text{user}}$ 。

[0054] 步骤 S2,对所述候选特征项进行频次过滤;具体的讲,就是在训练语料中统计候选特征项的出现次数,以频次作为指标进行过滤,对频次大于等于阈值的候选特征项保留,小于阈值的候选特征项剔除,可以调整阈值对保留的范围进行控制。

[0055] 步骤 S3,对所述候选特征项进行频次再过滤;具体的过滤过程包括:

[0056] 首先,对不合理的频次进行重新估计,比如,若所有出现 B 时都是 AB 的情况,则 B 的频次变为零。频次重新估计公式为:

[0057]

$$\begin{cases} \log_2|a| * f(a) & \text{当 } a \text{ 没有被包含现象;} \\ \log_2|w| * (f(a) - \frac{\sum_{b \in T_a} f(b)}{P(T_a)}) & \text{其它;} \end{cases}$$

[0058] 其中,a 表示特征项;f(a) 表示 a 的词频;b 表示包含了 a 的长串特征项; T_a 表示 b 的集合; $P(T_a)$ 表示集合大小。

[0059] 然后,以重新评估后的频次作为指标进行再次过滤,对频次大于等于阈值的候选特征项保留,小于阈值的候选特征项剔除,可以调整阈值,对保留的范围进行控制。

[0060] 步骤 S4 :对所述候选特征项进行自动选择,从而提取特征项。具体的讲,就是该步骤将正向语料从所述步骤 S3 中获取到的候选特征项与负向语料从所述步骤 S3 中获取的候选特征项进行合并,因此合并后这些候选特征项有两个词频,分别对应正向频次和负向频次。采用统计学的卡方统计量来进行特征项的自动选择,保留卡方值最大的前 N 个候选特征项作为最终特征项信息。卡方统计量公式为:

[0061]

$$\chi^2(w_i, C_k) = \frac{N(AD - BC)^2}{(A + C)(A + B)(B + D)(C + D)}$$

[0062] 其中 A、B、C、D、N 的含义如下:

	属于 C_k 的文 本	不属于 C_k 的文 本	总 数
包含 w 特征项	A	B	A+B
不包含 w 特征 项	C	D	C+D
总数	A+C	B+D	N

[0064] 表中 k 只取 0 或 1,代表两种类别,即正向类别和负向类别。

[0065] 需要说明的是,所述特征项包括单字词和多字词。单字词对负向文本的判定影响较大。特别是论坛文本信息的内容,单字的切分单元较常见,如果不考虑单字,对负向文本很容易造成误判。

[0066] 2043 :统计所述特征项的语料信息得分;在步骤 S4 中已保存了所述特征项的频次,且每个特征项都有两个频次,分别代表正向频次和负向频次,比如,“发票”的正向频次要远远大于负向频次,因为“发票”更常见于广告的不良信息中。将每个特征项的正向频次看作是特征项的正向权重,将每个特征项的负向频次看作是特征项的负向权重。对于所有特征项,分别对正负向权重进行归一化,这样,权重值才有比较意义。归一化的公式为:

$$score(w_i) = \frac{freq(w_i)}{\sum freq(w_i)}$$

[0068] 由于生成的特征项及其权重是根据系统预先准备的标准两类语料训练得到的,保存生成结果作为系统预研模型特征项信息。

[0069] 将所述预处理后的待过滤文本信息与所述系统预研模型特征项信息进行特征信息匹配,获得待过滤文本特征项信息,计算所述特征项信息正向得分,其计算公式为:

[0070]

$$\text{score}_{\text{pos}}(\text{doc}) = \sum \log(\text{score}(w_i)_{\text{pos}})$$

[0071] 计算所述特征项信息负向得分,其计算公式为:

[0072]

$$\text{score}_{\text{neg}}(\text{doc}) = \sum \log(\text{score}(w_i)_{\text{neg}})$$

[0073] 同时,考虑到 num(ad) 与 num(user),上述计算公式右侧变化为:

[0074]

$$\sum \log(\text{score}(w_i)_{\text{neg}}) + \text{num}(\text{ad}) * \text{score}_{\text{ad}} + \text{num}(\text{user}) * \text{score}_{\text{user}}$$

[0075] 2044:根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;若 $\text{score}_{\text{pos}}(\text{doc}) > \text{score}_{\text{neg}}(\text{doc})$,则系统预研模型信息判定此待处理文本为不良文本;若 $\text{score}_{\text{pos}}(\text{doc}) == \text{score}_{\text{neg}}(\text{doc})$,则此模型失效,判定失败;若 $\text{score}_{\text{pos}}(\text{doc}) < \text{score}_{\text{neg}}(\text{doc})$,则系统预研模型信息判定此待处理文本为正常文本。

[0076] 2045:根据判断结果,给出所述第一匹配结果。

[0077] 205:将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;该步骤具体可以包括的流程与步骤 204 所述流程大致相同。

[0078] 需要说明的是,所述获取用户反馈模型信息的过程与获取系统预研模型信息的过程主要不同的地方是步骤 201 中训练语料的选择。所述用户反馈模型信息的训练语料的来源还可以包括如下两方面:

[0079] (1) 用户反馈机制。用户在实际体验过程中发现判定出现问题的信息,主要是将不良信息判定为正常信息的情况,对系统进行报错,系统接收用户标准答案作为反馈语料。

[0080] (2) 判定模型机制。待处理文本进入步骤 206 的不良信息判定流程,输出对该文本的判定结果。结果包括的两种情况,即不良文本或者正常文本。根据判定可信度情况决定待处理文本是否参与反馈训练。

[0081] 206:根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本信息进行过滤处理。具体的讲,就是判断所述第一匹配结果与所述第二匹配结果的判定结果是否一致,即系统预研模型信息与用户反馈模型信息的判定结果。若判定相同,同为不良信息文本或正常信息文本,则判定结果可信度较大,可用于反馈训练;若判定不同,则判定结果可信度有损失,但若采取较为严格的过滤策略,则过滤此文本,但不可用于反馈训练;若其中有一模型失效,则结果依据剩余模型的判定结果,且认为有一定可信度,可用于反馈训练;若两个模型皆失效,则返回失效标志,不可用于反馈训练。

[0082] 需要注意的是,每完成一个待过滤文本信息的判定过程后,该方法还可以包括:

[0083] 获取所述用户反馈模型信息的语料数量以及其对应的阈值;具体的讲,就是统计可以用于反馈训练的语料数量,判断所述语料数量是否超出其对应阈值。

[0084] 根据所述用户反馈模型信息的语料数量以及其对应的阈值,对所述用户反馈模型信息进行更新。若语料数量大于阈值,则对反馈语料进行重新训练,更新用户反馈模型信息。调整阈值的大小,可以调整更新周期。

- [0085] 如图 3 所示,为本发明实施例提供的一种网络不良信息的过滤装置;该装置包括:
- [0086] 信息获取单元 301,用于获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息;
- [0087] 预处理单元 302,用于对所述待过滤文本信息进行预处理;
- [0088] 第一匹配单元 303,用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果;
- [0089] 第二匹配单元 304,用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;
- [0090] 过滤单元 305,用于根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本进行过滤处理。
- [0091] 如图 4 所示,为本发明实施例提供的一种网络不良信息的过滤装置;该装置包括:
- [0092] 信息获取单元 401,用于获取待过滤文本、系统预研模型信息以及用户反馈模型信息;还用于获取所述用户反馈模型信息的训练语料。其中,所述用户反馈模型信息的语料包括:用户反馈语料和/或被过滤语料。
- [0093] 预处理单元 402,用于对所述待过滤文本信息进行预处理;该单元具体包括:
- [0094] 切分子单元 4021,用于对所述待过滤文本信息进行切分处理;
- [0095] 统计子单元 4022,用于统计所述切分处理后的候选特征项数量。
- [0096] 第一匹配单元 403,用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果;该单元具体可以包括:
- [0097] 信息获取子单元 4031,用于获取所述预处理后的待过滤文本信息以及所述系统预研模型信息;其中,所述系统预研模型信息包括:规则索引库和所述系统预研模型特征项信息;
- [0098] 匹配子单元 4032,用于将所述预处理后的待过滤文本信息与所述系统预研模型信息进行匹配,获取特征项;
- [0099] 统计子单元 4033,用于统计所述特征项的语料信息得分;
- [0100] 判断子单元 4034,用于根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;
- [0101] 结果输出子单元 4035,用于根据判断结果,给出所述第一匹配结果。
- [0102] 第二匹配单元 404,用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;该单元具体可以包括:
- [0103] 信息获取子单元 4041,用于获取所述预处理后的待过滤文本信息以及所述用户反馈模型信息;其中,所述用户反馈模型信息包括:规则索引库和所述用户反馈模型特征项信息;
- [0104] 匹配子单元 4042,用于将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行匹配,获取特征项;
- [0105] 统计子单元 4043,用于统计所述特征项的语料信息得分;
- [0106] 判断子单元 4044,用于根据所述语料信息得分,判断所述特征项所对应的待过滤文本信息是否为不良信息;
- [0107] 结果输出子单元 4045,用于根据判断结果,给出所述第二匹配结果。

[0108] 过滤单元 405,用于根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本信息进行过滤处理。

[0109] 需要注意的是,该装置还包括:

[0110] 阈值获取单元 406,用于获取所述用户反馈模型信息的语料数量以及其对应的阈值;

[0111] 更新单元 407,用于根据所述用户反馈模型信息的语料数量以及其对应的阈值,对所述用户反馈模型信息进行更新。当所述阈值获取单元获取到的用户反馈模型信息的语料数量达到其对应的阈值时,所述更新单元根据所述用户反馈模型信息的语料数量以及其对应的阈值,对所述用户反馈模型信息进行更新。

[0112] 本发明实施例提供的网络不良信息的过滤方法以及装置,通过获取待过滤文本信息、系统预研模型信息以及用户反馈模型信息;对所述待过滤文本信息进行预处理;将所述预处理后的待过滤文本信息与所述系统预研模型信息进行特征信息匹配,给出第一匹配结果;将所述预处理后的待过滤文本信息与所述用户反馈模型信息进行特征信息匹配,给出第二匹配结果;根据所述第一匹配结果与所述第二匹配结果,对所述待过滤文本信息进行过滤处理。由于本发明中采用了两次匹配进行系统过滤,所以系统自动过滤不良信息的准确性较高,从而提高了系统的性能;还由于本发明实施例采用了用户反馈模型信息进行不良信息的过滤,使得用户反馈信息能够及时的应用于系统自动过滤不良信息的过程中,从而实现了系统的匹配信息自动更新的功能。

[0113] 通过以上的实施方式的描述,本领域普通技术人员可以理解:实现上述实施例方法中的全部或部分步骤是可以通程序来指令相关的硬件来完成,所述的程序可以存储于一计算机可读取存储介质中,该程序在执行时,包括如上述方法实施例的步骤,所述的存储介质,如:ROM/RAM、磁碟、光盘等。

[0114] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

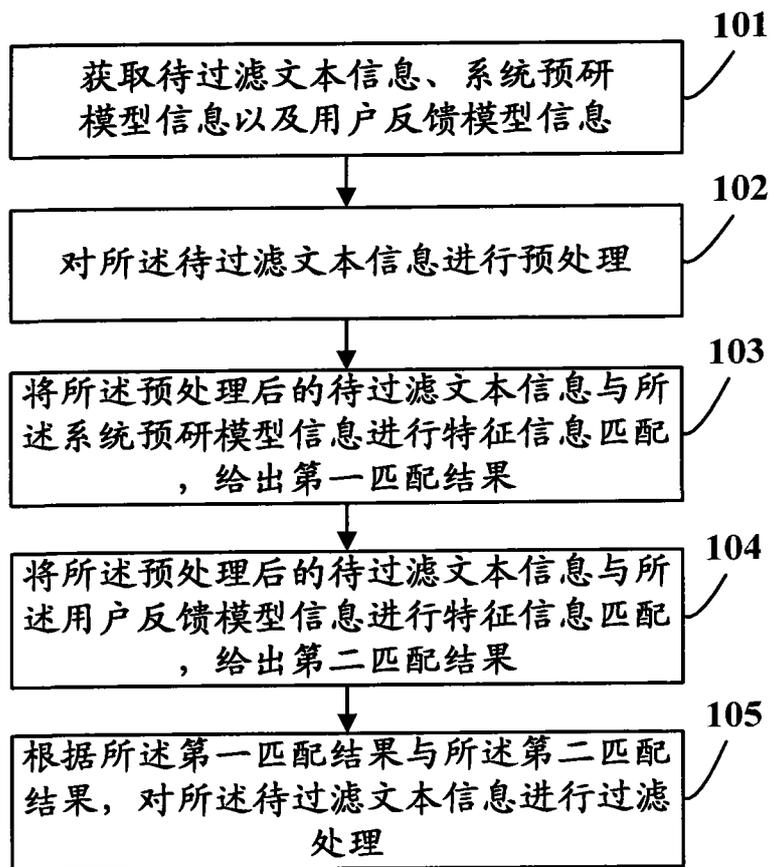


图 1

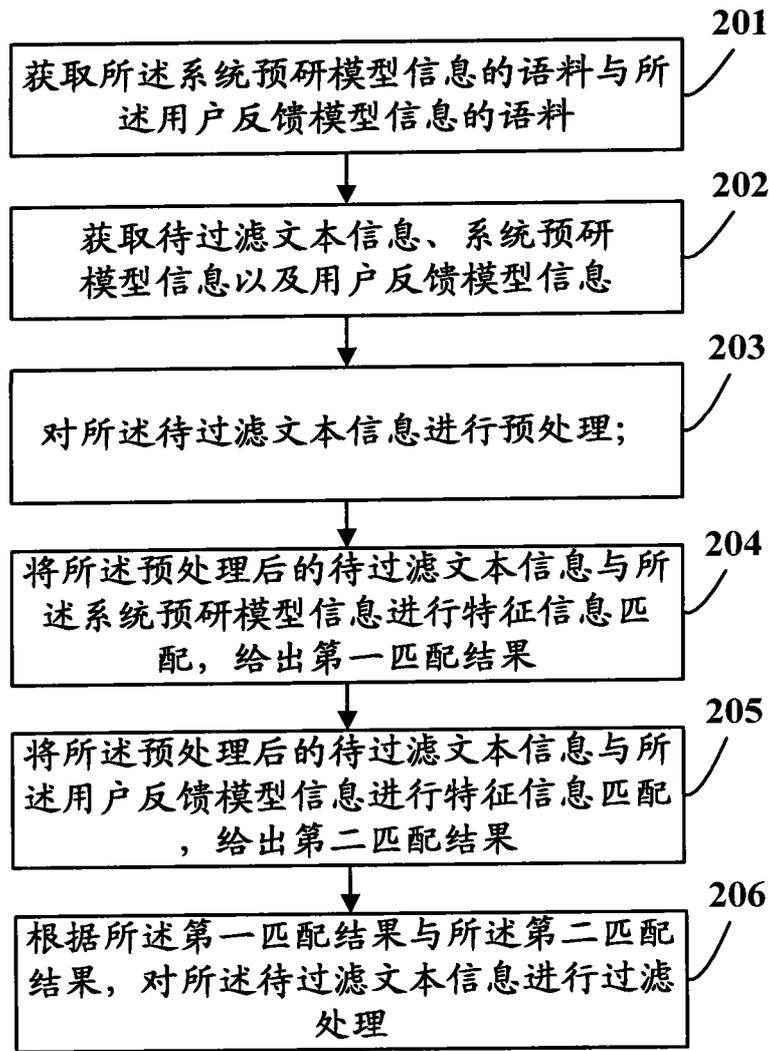


图 2

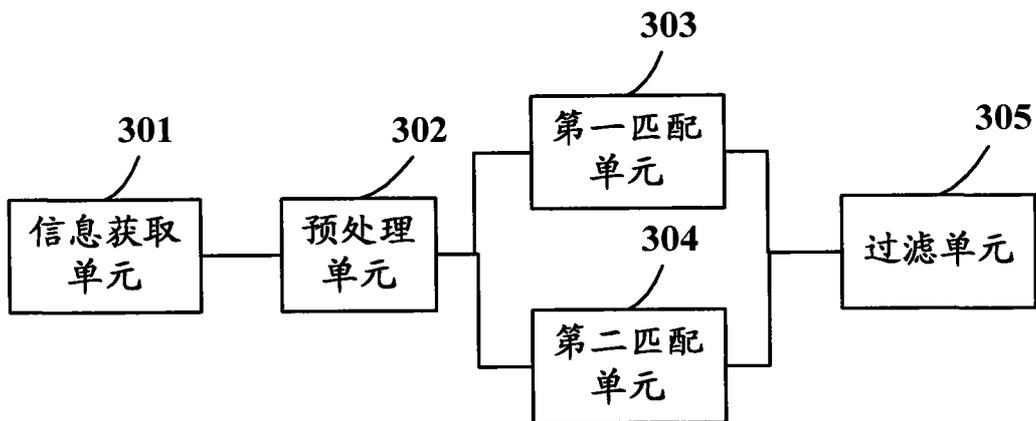


图 3

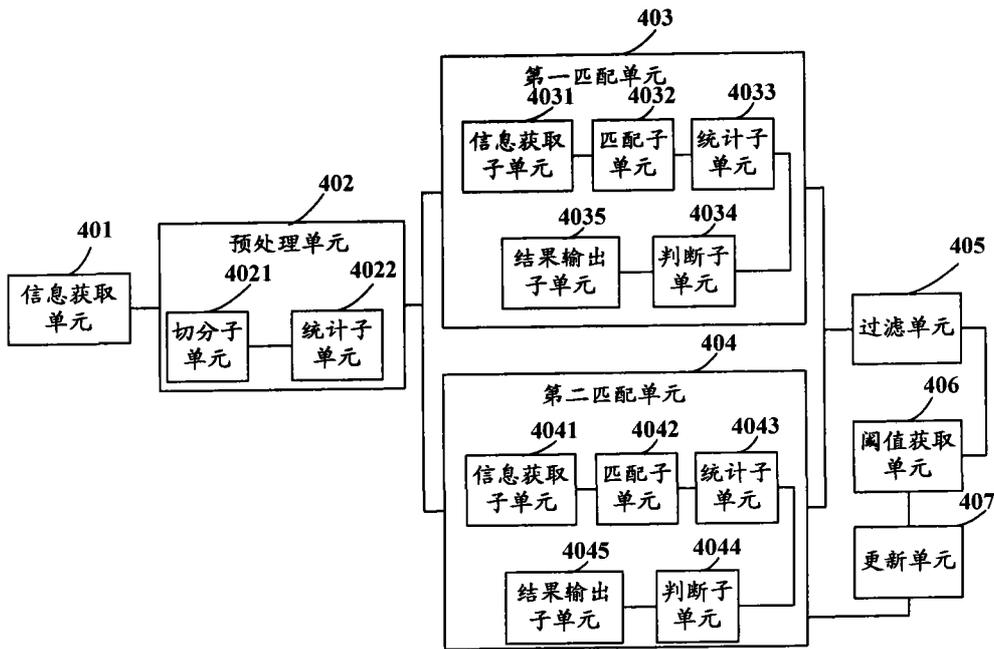


图 4