(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 105468677 A
(43)申请公布日 2016.04.06

(21)申请号 201510781245.7

(22)申请日 2015.11.13

(71)申请人 国家计算机网络与信息安全管理中心
地址 100029 北京市朝阳区裕民路甲3号

(72)发明人 吕雁飞 王树鹏 张鸿 丁煜 樊冬进 肖东方 郑亚松 周晓阳 何慧虹 史亮

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

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

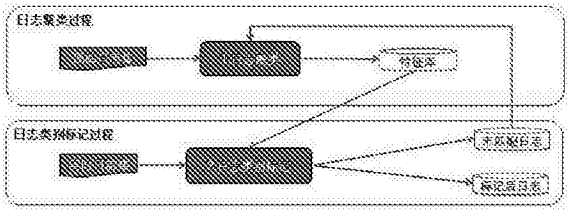
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于图结构的日志聚类方法

(57)摘要

本发明提供一种基于图结构的日志聚类方法，包括，基于文本分词、向量相似度以及最大连通子图对日志进行聚类，获取特征库；并根据特征库中的类别特征对海量日志进行类别标记；该方法无需人工指定聚类数目，自动识别海量日志中最合适的类别数目；另外，该方法可对日志进行精准分类，为海量日志数据挖掘奠定了基础。



CN 105468677 A

1.一种基于图结构的日志聚类方法，其特征在于，所述方法包括：基于文本分词、向量相似度以及对最大连通子图日志进行聚类，获取特征库；并根据特征库中的类别特征对海量日志进行类别标记。

2.根据权利要求1所述的方法，其特征在于，所述获取特征库包括下述步骤：

(1)将原始日志结构化，生成结构化日志数据；

(2)构建分词库；

(3)依据词库将结构化的日志向量化；

(4)删除重复的日志向量；

(5)确定日志相似关系图，生成各个类别包含的日志向量集合；

(6)构建特征库。

3.根据权利要求2所述的方法，其特征在于，所述步骤(1)中，生成结构化日志数据包括：输入原始日志，对半结构化的原始日志按列结构化，输出结构化日志数据。

4.根据权利要求2所述的方法，其特征在于，所述步骤(2)中，构建分词库包括，输入结构化日志数据，获取所述结构化日志数据包含的所有分词，并依据预设规则删除干扰词，生成日志数据分词库，该分词库中每个分词对应一个编号；其中，

所述干扰词，包括IP地址、端口号和16进制数字。

5.根据权利要求2所述的方法，其特征在于，所述步骤(3)中，日志向量化包括，将包含日志核心内容的字段进行分词，将获取的分词与词库相匹配，并用词库中分词编号代替分词，忽略未包含在词库中的分词，并保持分词原有的相对顺序，将文本转化为向量。

6.根据权利要求2所述的方法，其特征在于，所述步骤(5)中，确定日志相似关系图包括：将去重后的日志向量映射为图中的一个点，计算点与点之间的相似度；

若两个日志向量相似，则所述日志向量之间存在一条边。

7.根据权利要求6所述的方法，其特征在于，所述判定相似度包括，设A和B分别表征两个日志向量$A=(a_1,a_2,\cdots,a_m)$，$B=(b_1,b_2,\cdots,b_n)$；LCS({A,B})为向量A和B的最长公共子序列；若该最长公共子序列的长度与日志向量A和B比值皆高于经验阈值TH，则为相似，其表达式为：

$$\frac{LCS.length}{|A|}\geq TH \quad and \quad \frac{LCS.length}{|B|}\geq TH \qquad (1)。$$

8.根据权利要求7所述的方法，其特征在于，所述步骤(5)中，生成各个类别包含的日志向量集合包括；

将日志相似关系图中的每个最大连通子图定义为一个类，每一类包含的日志向量即该最大连通子图包含的点。

9.根据权利要求2所述的方法，其特征在于，所述步骤(6)中，构建特征库包括：各个日志类别的特征为该类别包含的所有日志向量的最长公共子序列；设第i类集合$R_i=\{S_1,S_2,\cdots,S_p\}$，$LCS(R_i)$为第i类中所有日志向量的最长公共子串，$w_i$为第i类的特征，其中$w_i=LCS(R_i)$；

输入每个日志类别所包含的日志向量集合，输出特征库。

10.根据权利要求1所述的方法，其特征在于，所述对海量日志进行类标记，具体步骤包括：

实时采集日志数据，将日志结构化，输出结构化日志数据；

对日志核心内容的字段进行分词，按预设规则去除干扰词；将日志分词集合中的每个词和原词库相匹配，若存在新词，则将该新词添加至词库，并输出新词库；

所述对日志进行结构化包括：

输入新词库和日志数据；

将日志数据由文本转为向量；

将包含日志内容的字段进行分词，将所述分词与词库匹配，用词库中分词的编号代替分词，忽略未包含词库中的分词，并保持分词原有的相对顺序，将文本转化为向量并输出；

所述日志类别匹配包括：

输入日志向量和通过日志聚类获得的特征库；

计算日志向量与特征库中各类别特征的相似度；若日志向量和特征w_i符合相似规则，则将该日志标记为第i类，输出携带标记的日志；

若日志与特征库中任意类别特征皆不相似，则匹配失败；将该日志存放于故障知识库，并定期重新进行聚类，生成新的类别特征，以更新特征库。

一种基于图结构的日志聚类方法

技术领域

[0001] 本发明涉及文本聚类领域，具体涉及一种基于图结构的日志聚类方法。

背景技术

[0002] 随着信息技术的飞速发展和集群规模的不断扩大，随之产生海量日志数据，然而却没有对日志数据进行有效的分析与挖掘。日志数据记录了系统的运行信息，挖掘日志数据具有重要意义，例如通过分析日志数据我们可以构建智能运维系统，完成故障定位、故障预警等功能。对日志进行精准的类别标记，是日志数据挖掘的重要方向。基于此我们通过对海量日志聚类，自动识别日志合适的类别数目。通过提取各类别特征，生成日志类别特征库，并根据特征库对新日志进行类别标记。其中，日志聚类方法的选择是重中之重。传统的聚类算法并不能完成海量日志聚类的需求。例如传统的K-Means、K-Medoid聚类算法，要求指定聚类的个数，不能自动识别日志合适的类别数目。传统的Denclue聚类算法为了得到较佳的聚类效果，需要经过不断的实验来获得合适的聚类数目，参数难以控制，计算量过大，且聚类不能保证得到真实的类别数目。因此需要探索新的日志聚类模型。

发明内容

[0003] 为克服上述缺陷，本发明提供一种基于图结构的日志聚类方法，依据类别特征知识库可对日志进行分类，大大提高了面对海量日志聚类的精确度。

[0004] 为了实现上述发明目的，本发明采取如下技术方案：

[0005] 一种基于图结构的日志聚类方法，所述方法包括：基于文本分词、向量相似度以及对最大连通子图日志进行聚类，获取特征库；并根据特征库中的类别特征对海量日志进行类别标记。

[0006] 优选的，所述获取特征库包括下述步骤：

[0007] (1)将原始日志结构化，生成结构化日志数据；

[0008] (2)构建分词库；

[0009] (3)依据词库将结构化的日志向量化；

[0010] (4)删除重复的日志向量；

[0011] (5)确定日志相似关系图，生成各个类别包含的日志向量集合；

[0012] (6)构建特征库。

[0013] 进一步地，所述步骤(1)中，生成结构化日志数据包括：输入原始日志，对半结构化的原始日志按列结构化，输出结构化日志数据。

[0014] 进一步地，所述步骤(2)中，构建分词库包括，输入结构化日志数据，获取所述结构化日志数据包含的所有分词，并依据预设规则删除干扰词，生成日志数据分词库，该分词库中每个分词对应一个编号；其中，

[0015] 所述干扰词，包括IP地址、端口号和16进制数字。

[0016] 进一步地，所述步骤(3)中，日志向量化包括，将包含日志核心内容的字段进行分

词，将获取的分词与词库相匹配，并用词库中分词编号代替分词，忽略未包含在词库中的分词，并保持分词原有的相对顺序，将文本转化为向量。

[0017] 进一步地，所述步骤(5)中，确定日志相似关系图包括：将去重后的日志向量映射为图中的一个点，计算点与点之间的相似度；

[0018] 若两个日志向量相似，则所述日志向量之间存在一条边。

[0019] 进一步地，所述判定相似度包括，设A和B分别表征两个日志向量$A=(a_1,a_2,\cdots,a_m)$，$B=(b_1,b_2,\cdots,b_n)$；LCS({A,B})为向量A和B的最长公共子序列；若该最长公共子序列的长度与日志向量A和B比值皆高于经验阈值TH，则为相似，其表达式为：

[0020] $$\frac{LCS.length}{|A|}\geq TH \quad and \quad \frac{LCS.length}{|B|}\geq TH \qquad (1)。$$

[0021] 进一步地，所述步骤(5)中，生成各个类别包含的日志向量集合包括；

[0022] 将日志相似关系图中的每个最大连通子图定义为一个类，每一类包含的日志向量即该最大连通子图包含的点。

[0023] 进一步地，所述步骤(6)中，构建特征库包括：各个日志类别的特征为该类别包含的所有日志向量的最长公共子序列；设第i类集合$R_i=\{S_1,S_2,\cdots,S_p\}$，$LCS(R_i)$为第i类中所有日志向量的最长公共子串，$w_i$为第i类的特征，其中$w_i=LCS(R_i)$；

[0024] 输入每个日志类别所包含的日志向量集合，输出特征库。

[0025] 优选的，所述对海量日志进行类标记，具体步骤包括：

[0026] 实时采集日志数据，将日志结构化，输出结构化日志数据；

[0027] 对日志核心内容的字段进行分词，按预设规则去除干扰词；将日志分词集合中的每个词和原词库相匹配，若存在新词，则将该新词添加至词库，并输出新词库；

[0028] 所述对日志进行结构化包括：

[0029] 输入新词库和日志数据；

[0030] 将日志数据由文本转为向量；

[0031] 将包含日志内容的字段进行分词，将所述分词与词库匹配，用词库中分词的编号代替分词，忽略未包含词库中的分词，并保持分词原有的相对顺序，将文本转化为向量并输出；

[0032] 所述日志类别匹配包括：

[0033] 输入日志向量和通过日志聚类获得的特征库；

[0034] 计算日志向量与特征库中各类别特征的相似度；若日志向量和特征w_i符合相似规则，则将该日志标记为第i类，输出携带标记的日志；

[0035] 若日志与特征库中任意类别特征皆不相似，则匹配失败；将该日志存放于故障知识库，并定期重新进行聚类，生成新的类别特征，以更新特征库。

[0036] 与最接近的现有技术相比，本发明达到的有益效果是：

[0037] 该日志聚类方法，一方面无需人工指定聚类数目，自动识别海量日志中最合适的类别数目，并支持大规模的日志聚类问题。另一方面，该方法有效保障了对原始的日志的类别标记精准度，完成了日志的精准分类，并支持日志海量日志数据的实时分类以及离线分类，为海量日志数据挖掘奠定了基础。

附图说明

[0038] 图1为一种基于图结构的日志聚类方法总流程图；

[0039] 图2为日志聚类方法流程图；

[0040] 图3为日志向量化结构示意图；

[0041] 图4为日志相似关系示意图；

[0042] 图5为日志类别标记方法流程图。

具体实施方式

[0043] 以下将结合附图，对本发明的具体实施方式作进一步的详细说明。

[0044] 如图1所示，一种基于图结构的日志聚类方法，所述方法包括：基于文本分词、向量相似度以及对最大连通子图日志进行聚类，获取特征库；并根据特征库中的类别特征对海量日志进行类别标记。

[0045] 1、获取特征库包括下述步骤：

[0046] (1)将原始日志结构化，生成结构化日志数据；包括：输入原始日志，对半结构化的原始日志按列结构化，输出结构化日志数据。

[0047] 例如Linux syslog日志形式如表1.1所示，按列结构化为Timestamp、Level、Source、Message等字段。原始syslog经结构化处理后变为表1.2中格式：

[0048] 表1.1

[0049]

Syslog 日志记录一	2014-12-18 22:00:19; snmpd[2247]:Connection from UDP: [255.255.255.255]: 62959
Syslog 日志记录二	2014-12-18 22:00:40; snmpd[2247]:Receiced SNMP packet(s) from UDP: [127.0.0.1]:58656

[0050] 表1.2

[0051]

Timestamp	Source	Message
2014-12-18 22:00:19	Snmpd[2447]	Connection from UDP 255.255.255.255 62959
2014-12-18 22:00:40	Snmpd[2447]	Receive SNMP packet(s) from UDP 127.0.0.1 58656

[0052] (2)构建分词库；包括：输入结构化日志数据，获取所述结构化日志数据包含的所有分词，并依据预设规则删除干扰词，生成日志数据分词库，该分词库中每个分词对应一个编号；其中，

[0053] 所述干扰词，包括IP地址、端口号和16进制数字。

[0054] 例如上表的Message字段中“255.255.255.255|62959|127.0.0.1|58656”等分词是日志影响日志聚类的干扰词，可选择按正则表达式定义规则将其去除，得到表1.3所示分词库：

[0055] 表1.3

[0056]

编号	Token
1	accepting
2	access
3	address

[0057]

4	after
5	again
6	and

[0058] (3)依据词库将结构化的日志向量化；

[0059] 将包含日志核心内容的字段进行分词，将获取的分词与词库相匹配，并用词库中分词编号代替分词，忽略未包含在词库中的分词，并保持分词原有的相对顺序，将文本转化为向量。例如：Connection from UDP:[255.255.255.255]:62959向量化为(35,65,181)，Received SNMP packet(s)from UDP:[127.0.0.1]:58656向量化为(147,168,133,161,65,181)。

[0060] 如图3所示，(4)删除重复的日志向量；日志向量中包含许多的相同向量；去除掉相同的向量，获得去除干扰词的无重复的日志向量集合。

[0061] (5)确定日志相似关系图，生成各个类别包含的日志向量集合；可以使用余弦相似性、最长公共子序列等。

[0062] 其中，确定日志相似关系图包括：将去重后的日志向量映射为图中的一个点，判定其相似度；若两个日志向量相似，则所述日志向量之间存在一条边。

[0063] 判定相似度：设A和B分别表征两个日志向量$A=(a_1,a_2,\cdots,a_m)$，$B=(b_1,b_2,\cdots,b_n)$；LCS({A,B})为向量A和B的最长公共子序列；若该最长公共子序列的长度与日志向量A和B比值皆高于经验阈值TH，则为相似，其表达式为：

[0064] $$\frac{LCS.length}{|A|}\geq TH \quad and \quad \frac{LCS.length}{|B|}\geq TH \qquad (1)。$$

[0065] 例如将两个日志向量的相似性度量可定义为：设A、B代表两个日志向量，其中$A=(a_1,a_2,\cdots,a_m)$，$B=(b_1,b_2,\cdots,b_n)$，LCS({A,B})表示A和B的最长公共子序列。如LCS({(1,2,1,2,3),(3,1,2,3,4)})=(1,2,3)表示日志向量(1,2,1,2,3)和(3,1,2,3,4)的最长公共子序列为(1,2,3)。如果最长公共子序列的长度与这两个日志向量长度的比值，都高于一个人工经验确定的阈值(TH)，即如公式(1)所示，则判定两个日志向量相似。

[0066] 生成各个类别包含的日志向量集合包括：将日志相似关系图中的每个最大连通子图定义为一个类，每一类包含的日志向量即该最大连通子图包含的点。如图4所示，图中包含4个最大连通子图，分别为{a,b,c},{g,h},{e,d,f},{i}，即日志向量集合包含4个类，分别为{a,b,c},{g,h},{e,d,f},{i}。

[0067] (6)构建特征库。包括：各个日志类别的特征为该类别包含的所有日志向量的最长公共子序列；设第i类集合$R_i=\{S_1,S_2,\cdots,S_p\}$，$LCS(R_i)$为第i类中所有日志向量的最长公共子串，$w_i$为第i类的特征，其中$w_i=LCS(R_i)$；输入每个日志类别所包含的日志向量集合，输出特征库。例如$w_i=LCS(\{(a,b,a,b,e),(c,a,b,e),(a,b,e,d)\})=(a,b,e)$，则表示第i类的特征为(a,b,e)。

[0068] 如图5所示，2、对海量日志进行类标记，具体步骤包括：

[0069] 实时采集日志数据，将日志结构化，输出结构化日志数据；

[0070] 将当前词库中未包含的新日志中未出的分词，添加至词库；

[0071] 对日志核心内容的字段进行分词，按预设规则去除干扰词；将日志分词集合中的每个词和原词库相匹配，若存在新词，则将该新词添加至词库，并输出新词库。

[0072] 对日志进行结构化包括：

[0073] 输入新词库和日志数据；

[0074] 将日志数据由文本转为向量；

[0075] 将包含日志内容的字段进行分词，将所述分词与词库匹配，用词库中分词的编号代替分词，忽略未包含词库中的分词，并保持分词原有的相对顺序，将文本转化为向量并输出。

[0076] 日志类别匹配包括：

[0077] 输入日志向量和通过日志聚类获得的特征库；

[0078] 计算日志向量与特征库中各类别特征的相似度；若日志向量和特征w_i符合相似规则，则将该日志标记为第i类，输出携带标记的日志；

[0079] 若日志与特征库中任意类别特征皆不相似，则匹配失败；将该日志存放于故障知识库，并定期重新进行聚类，生成新的类别特征，以更新特征库。

[0080] 最后应当说明的是：以上实施例仅用以说明本发明的技术方案而非对其限制，所属领域的普通技术人员参照上述实施例依然可以对本发明的具体实施方式进行修改或者等同替换，这些未脱离本发明精神和范围的任何修改或者等同替换，均在申请待批的本发明的权利要求保护范围之内。

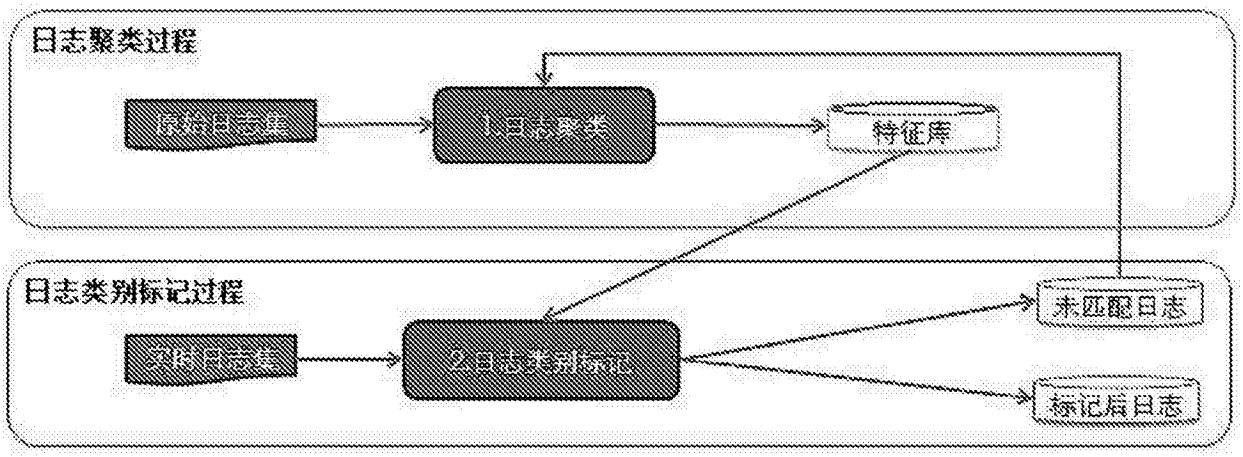

日志聚类过程
原始日志集
1.日志聚类
特征库
日志类别标记过程
待标日志集
2.日志类别标记
未匹配日志
标记后日志

图1

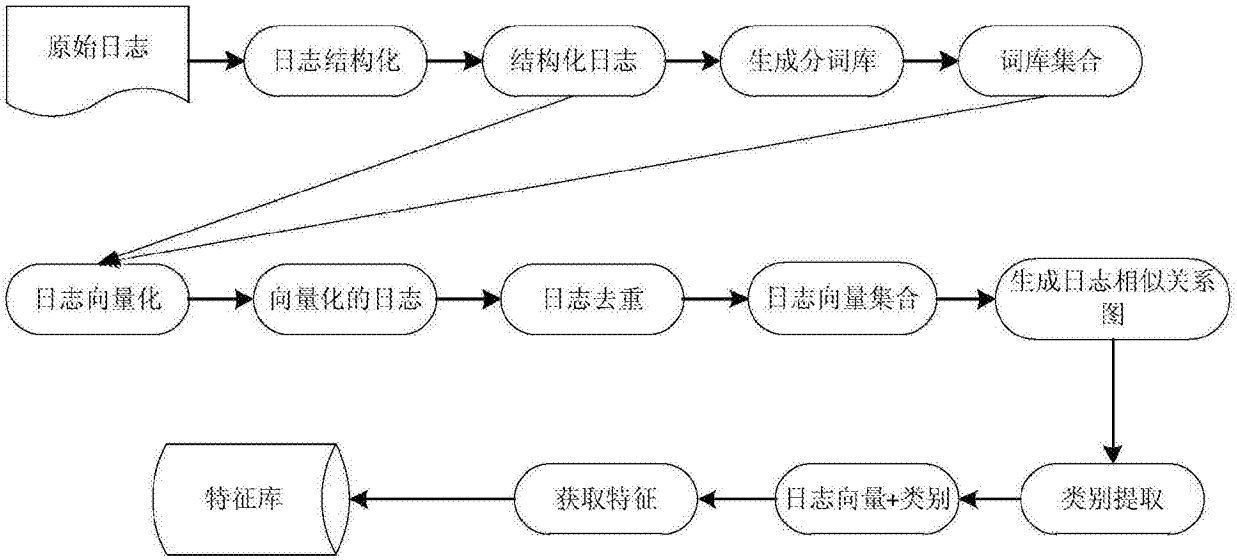

原始日志
日志结构化
结构化日志
生成分词库
词库集合
日志向量化
向量化的日志
日志去重
日志向量集合
生成日志相似关系图
特征库
获取特征
日志向量+类别
类别提取

图2

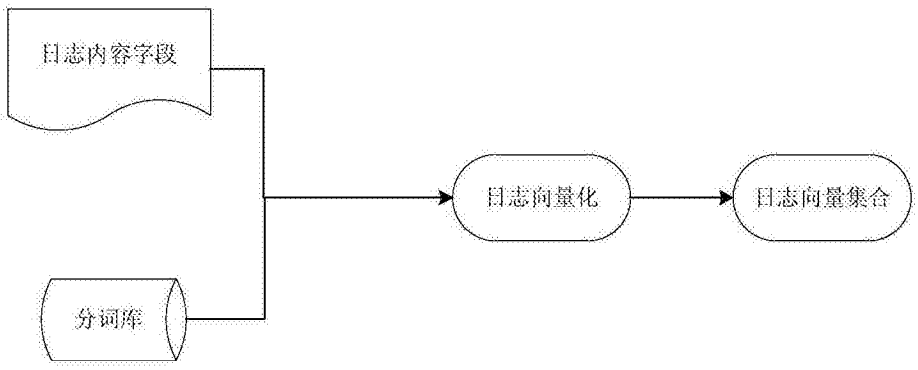
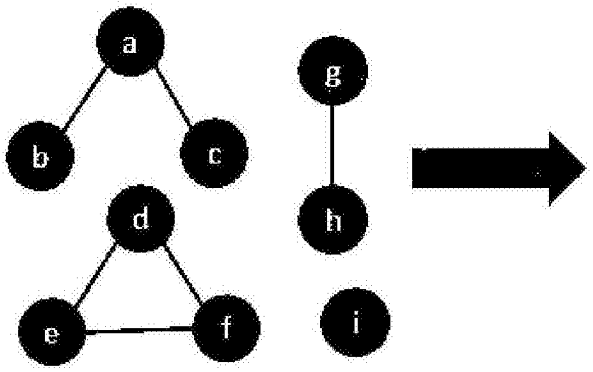

日志内容字段
日志向量化
日志向量集合
分词库

图3



类别	单个类集合
Label_1	a、b、c
Label_2	d、e、f
Label_3	g、h
Label_4	i

图4

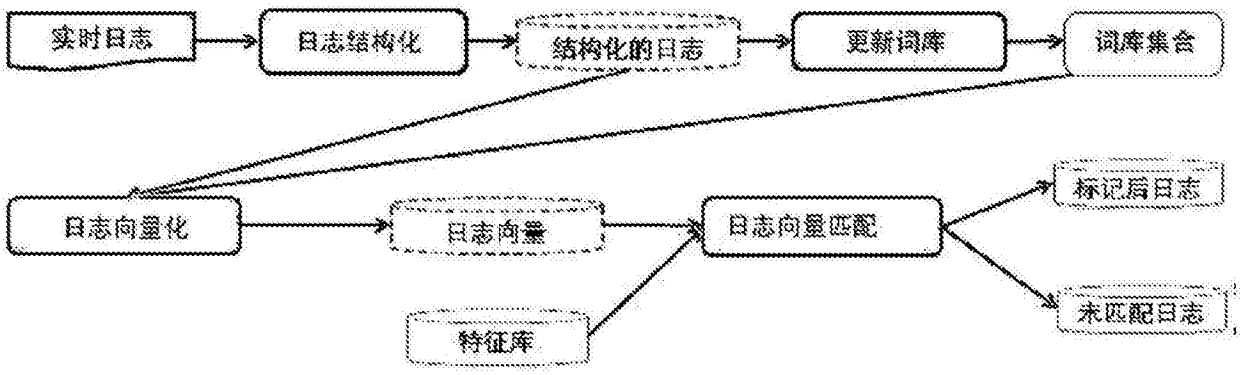


图5