(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 107169020 A
(43)申请公布日 2017.09.15

(21)申请号 201710224022.X

(22)申请日 2017.04.07

(71)申请人 南京邮电大学
地址 210023 江苏省南京市栖霞区文苑路9号

(72)发明人 徐小龙 杨春春

(74)专利代理机构 南京经纬专利商标代理有限公司 32200
代理人 田凌涛

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

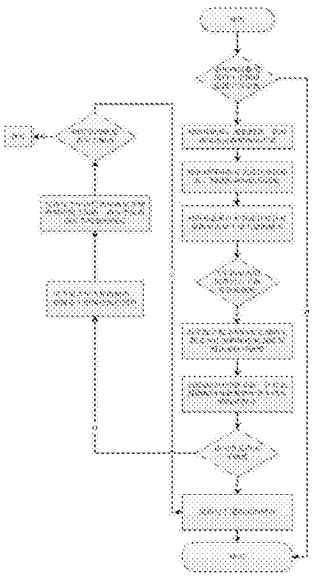
权利要求书2页 说明书8页 附图2页

(54)发明名称

一种基于关键字的定向网页采集方法

(57)摘要

本发明涉及一种基于关键字的定向网页采集方法，引入了文本加权算法为关键词设置权重，结合空间向量模型算法计算网页主题相关度，并且利用网页链接结构与主题相关度来评判网页的重要性。根据文本聚类算法将相关主题网页文档聚集在一起，利用朴素贝叶斯算法计算出待抓取的网页属于主题文档的概率。设置适应度函数筛选与主题相关的网页，依据网页的实时抓取情况动态的调整系统模型。本文基于分布式平台，结合开源网络采集架构，利用自适应主题算法实现对主题网页的定向抓取。采用分布式技术实现并行化抓取网页，充分利用各个节点计算资源，提高了网页的抓取速率。



CN 107169020 A

1.一种基于关键字的定向网页采集方法，用于在指定网页数据集中，查找与指定主题关键字相关的网页，其特征在于，包括如下步骤：

步骤1.判断指定网页数据集中的网页个数是否小于预设最大待采集爬行页面数，是则进入步骤6，否则进入步骤2；

步骤2.在指定网页数据集中随机提取最大待采集爬行页面数量的网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤4；

步骤3.分别针对各个待处理网页，分别执行如下步骤301至步骤302，获得各个待处理网页分别与指定主题关键字的相关概率；

步骤301.针对待处理网页的正文进行分词操作，构建该待处理网页正文所对应的分词集；

步骤302.根据该待处理网页正文所对应分词集中的各个分词，采用朴素贝叶斯算法计算获得该待处理网页与指定主题关键字的相关概率；

步骤4.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤5；

步骤5.将该指定主题关键字，以及该指定主题关键字所相关的采集网页集合作为训练样本，训练获得该指定主题关键字所对应的网页相关度采集器，通过该网页相关度采集器实现与该指定主题关键字相关网页的查找，所设计定向网页采集方法结束；

步骤6.根据预设适应度评价函数，计算获得指定网页数据集中各个网页的链接得分，并选取链接得分大于预设链接得分阈值的各个网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤7；

步骤7.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤8；

步骤8.选取步骤6中，将链接得分不大于预设链接得分阈值的各个网页，作为初级待处理网页，并针对所有初级待处理网页进行交叉变异操作，获得各个中级待处理网页，然后进入步骤9；

步骤9.获得各个中级待处理网页分别所对应的父页面，将各个父页面作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，并将与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，加入到该指定主题关键字所相关的采集网页集合中，并返回步骤5。

2.根据权利要求1所述基于关键字的定向网页采集方法，其特征在于：所述步骤3中还包括步骤301-1如下，执行完步骤301之后，进入步骤301-1，执行完步骤301-1之后，进入步骤302；

步骤301-1.采用tf-idf算法计算待处理网页正文所对应分词集中各个分词的权重，并根据各个分词的权重，针对该待处理网页正文所对应的分词集进行降维，更新该待处理网页正文所对应的分词集。

3.根据权利要求2所述基于关键字的定向网页采集方法，其特征在于：所述步骤3中还包括步骤301-2和步骤301-3如下，执行完步骤301-1之后，进入步骤301-2；执行完步骤301-2之后，进入步骤301-3；执行完步骤301-3之后，进入步骤302；

步骤301-2.采用空间向量模型算法，计算待处理网页正文所对应分词集中各个分词彼

此之间的相似度；

步骤301-3.针对该待处理网页正文所对应分词集中各个分词，根据各个分词彼此之间的相似度，采用k-means文本聚类算法，针对彼此之间相似度大于预设相似度阈值的各个分词进行聚合，更新该待处理网页正文所对应的分词集。

4.根据权利要求1所述基于关键字的定向网页采集方法，其特征在于：所述步骤6中的预设适应度评价函数如下：

$Fitness(link_i)=f_{sim}+f_{link}+f_{parent}+f_{datastruts}+f_{relevanturls/totalurls}+\lambda$

其中$Fitness(link_i)$代表第i个网页的连接得分；f_{sim}代表的是第i个网页预测主题相关度；f_{link}表示的是第i个网页URL的链接分析值；f_{parent}代表的是第i个网页的父页面的相关度，$f_{datastruts}$代表的是第i个网页URL的标签权重值，$f_{relevanturls/totalurls}$代表的是与指定主题关键字相关的网页数量与网页总数量的比值；λ是自适应调整的动态值。

一种基于关键字的定向网页采集方法

技术领域

[0001] 本发明涉及一种基于关键字的定向网页采集方法，属于主题网络爬虫、分布式计算的交叉技术领域。

背景技术

[0002] 随着电子计算机、存储设备、移动通信网络等信息技术的快速发展，移动互联网、社交网络、物联网等迅速普及，导致互联网平台数据量的迅速增长，大数据时代已经来临。据统计，截止2016年3月中旬，全球范围内，仅互联网上可知网页(不含隐含网页)总数已经超过46亿个，如何对网络数据的高效采集就显得尤为重要。

[0003] 数据采集是后续的数据挖掘、分析与决策的前提，网络数据采集的抓取效率直接决定着数据处理的效果，因此高效、精准的采集与主题相关的数据已经成为研究热点。高性能的主题网络采集必须考虑两方面：一方面是系统具有精准的定向主题采集策略，尽可能多的抓取与主题相关的网页，减少与主题无关的网页的采集；另一方面系统具有高度优化的整体架构性，方便管理并且具有高效的可扩展性，能够实现分布式的网页采集。

[0004] 随着数据量的增大，导致数据的采集和处理均需要实现分布式并行化，采用分布式技术可以有效提高数据抓取的速率。一种典型的分布式数据采集及处理平台，利用并行处理数据运算，产生了一种基于云平台的完全分布式平台，平台能够实现分布式网络数据的抓取、索引以及检索。像百度和谷歌等流行的很多商业通用搜索引擎，查询结果通常都是考虑广泛性而忽略了针对性，与面向特定领域的主题搜索引擎比较的话，专业针对性较弱，对采集结果的过滤和后期排序的相关度还有待提高。

[0005] 在传统的面目前的网络数据定向采集技术主要存在以下问题：(1)主题爬虫在保存网页之前需要对页面相关性进行判别，只保存与主题相关的网页。目前主题判别方法多是基于分类器的方法，分类器的准确率和效率都很低，难以实现对主题网页的准确抓取。(2)基于链接结构的主题采集算法主要问题是计算出的链接价值与主题的相关性较小，容易造成“主题漂移”，采集与主题无关的网页，并且基于链接内容评价的主题搜索策略搜索效率偏低。(3)目前的分布式定向数据采集系统，节点与节点之间需要进行频繁通信，并且系统的可扩展性不高。

发明内容

[0006] 本发明所要解决的技术问题是提供一种基于关键字的定向网页采集方法，不仅提高了采集主题网页的准确率，而且具有系统可扩展性强的优点，能够适用于分布式大规模网页的快速采集，而且采集方法能够较好的适用于分布式主题的环境下，并且算法的复杂度较低。

[0007] 本发明为了解决上述技术问题采用以下技术方案：本发明设计了一种基于关键字的定向网页采集方法，用于在指定网页数据集中，查找与指定主题关键字相关的网页，其特征在于，包括如下步骤：

[0008] 步骤1.判断指定网页数据集中的网页个数是否小于预设最大待采集爬行页面数，是则进入步骤6，否则进入步骤2；

[0009] 步骤2.在指定网页数据集中随机提取最大待采集爬行页面数量的网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤4；

[0010] 步骤3.分别针对各个待处理网页，分别执行如下步骤301至步骤302，获得各个待处理网页分别与指定主题关键字的相关概率；

[0011] 步骤301.针对待处理网页的正文进行分词操作，构建该待处理网页正文所对应的分词集；

[0012] 步骤302.根据该待处理网页正文所对应分词集中的各个分词，采用朴素贝叶斯算法计算获得该待处理网页与指定主题关键字的相关概率；

[0013] 步骤4.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤5；

[0014] 步骤5.将该指定主题关键字，以及该指定主题关键字所相关的采集网页集合作为训练样本，训练获得该指定主题关键字所对应的网页相关度采集器，通过该网页相关度采集器实现与该指定主题关键字相关网页的查找，所设计定向网页采集方法结束；

[0015] 步骤6.根据预设适应度评价函数，计算获得指定网页数据集中各个网页的链接得分，并选取链接得分大于预设链接得分阈值的各个网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤7；

[0016] 步骤7.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤8；

[0017] 步骤8.选取步骤6中，将链接得分不大于预设链接得分阈值的各个网页，作为初级待处理网页，并针对所有初级待处理网页进行交叉变异操作，获得各个中级待处理网页，然后进入步骤9；

[0018] 步骤9.获得各个中级待处理网页分别所对应的父页面，将各个父页面作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，并将与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，加入到该指定主题关键字所相关的采集网页集合中，并返回步骤5。

[0019] 作为本发明的一种优选技术方案：所述步骤3中还包括步骤301-1如下，执行完步骤 301之后，进入步骤301-1，执行完步骤301-1之后，进入步骤302；

[0020] 步骤301-1.采用tf-idf算法计算待处理网页正文所对应分词集中各个分词的权重，并根据各个分词的权重，针对该待处理网页正文所对应的分词集进行降维，更新该待处理网页正文所对应的分词集。

[0021] 作为本发明的一种优选技术方案：所述步骤3中还包括步骤301-2和步骤301-3如下，执行完步骤301-1之后，进入步骤301-2；执行完步骤301-2之后，进入步骤301-3；执行完步骤301-3之后，进入步骤302；

[0022] 步骤301-2.采用空间向量模型算法，计算待处理网页正文所对应分词集中各个分词彼此之间的相似度；

[0023] 步骤301-3.针对该待处理网页正文所对应分词集中各个分词，根据各个分词彼此

之间的相似度，采用k-means文本聚类算法，针对彼此之间相似度大于预设相似度阈值的各个分词进行聚合，更新该待处理网页正文所对应的分词集。

[0024] 作为本发明的一种优选技术方案，所述步骤6中的预设适应度评价函数如下：

[0025] $Fitness(link_i)=f_{sim}+f_{link}+f_{parent}+f_{datastruts}+f_{relevanturls/totalurls}+\lambda$

[0026] 其中$Fitness(link_i)$代表第i个网页的连接得分；f_{sim}代表的是第i个网页预测主题相关度；f_{link}表示的是第i个网页URL的链接分析值；f_{parent}代表的是第i个网页的父页面的相关度，$f_{datastruts}$代表的是第i个网页URL的标签权重值，$f_{relevanturls/totalurls}$代表的是与指定主题关键字相关的网页数量与网页总数量的比值；λ是自适应调整的动态值。

[0027] 本发明所述一种基于关键字的定向网页采集方法采用以上技术方案与现有技术相比，具有以下技术效果：

[0028] (1)本发明设计的基于关键字的定向网页采集方法，相较于传统方法，传统基于单机的定向主题采集策略需要消耗大量的时间和带宽，本发明的基于分布式的主题采集策略，利用分布式的方式让多台机器同时对网页进行采集，通过多个节点的并行的定向抓取网络数据，有效提高了数据采集到速率，缩短的采集的时间；

[0029] (2)本发明设计的基于关键字的定向网页采集方法中，采集数据准确性是判断采集系统优劣的重要指标之一，本发明的自适应主题采集算法所采集的数据具有较高的数据准确性，能够在降低系统开销的前提下，较为准确地采集与主题相关的网页；

[0030] (3)本发明设计的基于关键字的定向网页采集方法中，相较于单纯地人为设定阈值更加合理，能够根据实际的定向采集的与主题网页变化情况自适应调整阈值。以历史采集数据为参考，动态地制定合适的阈值，及时调整系统采集模型，从而实现又好又快的抓取。并且能够在一定程度上提高全局搜索性，避免了采集网页陷入局部最优的状态，通过自适应算法提高系统的整体采集准确率以便于合理地衡量新采集数据的变化程度。

附图说明

[0031] 图1是本发明所设计基于关键字的定向网页采集方法的示意图；

[0032] 图2是本发明所设计基于关键字的定向网页采集方法的分布式架构示意图。

具体实施方式

[0033] 下面结合说明书附图对本发明的具体实施方式作进一步详细的说明。

[0034] 本发明所设计基于关键字的定向网页采集方法，(1)为解决主题采集中的采集准确率不高，本文通过提出一种数据定向采集方法，以历史采集数据为参考，动态地制定合适的阈值，及时调整系统采集模型，从而实现又好又快的抓取。并且能够在一定程度上提高全局搜索性，避免了采集网页陷入局部最优的状态，通过自适应算法提高系统的整体采集准确率。(2)本文基于分布式平台，对分布式配置环境进行优化，利用Nutch开源爬虫框架，实现了自适应主题爬虫分布式多线程的对网页抓取。通过多个节点的并行的定向抓取网络数据，有效提高了数据采集到速率。具体而言，本发明采用以下技术方案解决上述技术问题。

[0035] 如图1和图2所示，本发明设计了一种基于关键字的定向网页采集方法，用于在指定网页数据集中，查找与指定主题关键字相关的网页，实际应用中，具体包括如下步骤：

[0036] 步骤1.判断指定网页数据集中的网页个数是否小于预设最大待采集爬行页面数，

是则进入步骤6，否则进入步骤2。

[0037] 步骤2.在指定网页数据集中随机提取最大待采集爬行页面数量的网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤4。

[0038] 步骤3.分别针对各个待处理网页，分别执行如下步骤301至步骤302，获得各个待处理网页分别与指定主题关键字的相关概率。

[0039] 步骤301.针对待处理网页的正文进行分词操作，构建该待处理网页正文所对应的分词集，然后进入步骤301-1。

[0040] 步骤301-1.采用tf-idf算法计算待处理网页正文所对应分词集中各个分词的权重，并根据各个分词的权重，针对该待处理网页正文所对应的分词集进行降维，更新该待处理网页正文所对应的分词集，然后进入步骤301-2。

[0041] 步骤301-2.采用空间向量模型算法，计算待处理网页正文所对应分词集中各个分词彼此之间的相似度，然后进入步骤301-3；

[0042] 步骤301-3.针对该待处理网页正文所对应分词集中各个分词，根据各个分词彼此之间的相似度，采用k-means文本聚类算法，针对彼此之间相似度大于预设相似度阈值的各个分词进行聚合，更新该待处理网页正文所对应的分词集，然后进入步骤302。

[0043] 步骤302.根据该待处理网页正文所对应分词集中的各个分词，采用朴素贝叶斯算法计算获得该待处理网页与指定主题关键字的相关概率。

[0044] 步骤4.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤5。

[0045] 步骤5.将该指定主题关键字，以及该指定主题关键字所相关的采集网页集合作为训练样本，训练获得该指定主题关键字所对应的网页相关度采集器，通过该网页相关度采集器实现与该指定主题关键字相关网页的查找，所设计定向网页采集方法结束。

[0046] 步骤6.根据如下预设适应度评价函数：

[0047] $Fitness(link_i)=f_{sim}+f_{link}+f_{parent}+f_{datastruts}+f_{relevanturls/totalurls}+\lambda$ (11)

[0048] 计算获得指定网页数据集中各个网页的链接得分，并选取链接得分大于预设链接得分阈值的各个网页，作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，然后进入步骤7。其中，$Fitness(link_i)$代表第i个网页的连接得分；f_{sim}代表的是第i个网页预测主题相关度；f_{link}表示的是第i个网页URL的链接分析值；f_{parent}代表的是第i个网页的父页面的相关度，$f_{datastruts}$代表的是第i个网页URL的标签权重值，$f_{relevanturls/totalurls}$代表的是与指定主题关键字相关的网页数量与网页总数量的比值；λ是自适应调整的动态值。

[0049] 步骤7.针对与指定主题关键字相关概率大于预设相关概率阈值的各个待处理网页，构建该指定主题关键字所相关的采集网页集合，并进入步骤8。

[0050] 步骤8.选取步骤6中，将链接得分不大于预设链接得分阈值的各个网页，作为初级待处理网页，并针对所有初级待处理网页进行交叉变异操作，获得各个中级待处理网页，然后进入步骤9。

[0051] 步骤9.获得各个中级待处理网页分别所对应的父页面，将各个父页面作为待处理网页，通过步骤3，获得各个待处理网页分别与指定主题关键字的相关概率，并将与指定主

题关键字相关概率大于预设相关概率阈值的各个待处理网页，加入到该指定主题关键字所相关的采集网页集合中，并返回步骤5。

[0052] 上述步骤3在具体实际应用中，具体如下：

[0053] 针对待处理网页的正文进行分词操作，构建该待处理网页正文所对应的分词集，通过文本加权算法对分词后的文本计算权重，把页面特征向量表示成为特征词的加权向量：

[0054] $V=\{V_1,V_2,V_3.....,V_n\}$ (1)

[0055] 式子(1)中，n表示文本特征的总数。v_i表示特征词t_i在向量V中加的权值，采用tf-idf 公式得到：

[0056] $V_i=tf_i*idf_i$ (2)

[0057] 式子(2)中，tf_i表示特征词t_i在页面中出现的频度，idf_i表示特征词t_i的倒文档频度，计算公式如(3)所示：tf_i表示某个词在文章出现的次数/文章的总词数：

[0058] $$tf_i=\frac{n_i}{\sum_{k=1}^{m}n_k} \quad (3)$$

[0059] 式子(3)中，n_i代表某个分词在文章出现的次数，n_k代表文章的总词数。

[0060] $$idf_i=\log_{10}\frac{\sum_{h=1}^{l}j_h}{b_i} \quad (4)$$

[0061] 式子(4)中，j_h表示页面样本集中的页面总数；b_i表示页面样本集中出现特征词t_i的页面的数量。

[0062] 将文档分词，经过关键词加权以后，选取文档中出现的关键词为该向量空间的一个向量，利用这些向量表示文档。文档中的关键词假设是一个r维空间向量，r代表关键词的个数。计算主题相关度的方法采用向量空间模型(VSM)算法，根据计算的结果对文档进行相似度的排序。但是网页是由许多标签组成的，不同的文档内容在不同的位置，代表的重要性也不一样；根据标签在网页中的位置设置不同的权值，如下表1所示。

[0063]

<title>	<meta>	<H><a>	其他
5.0	3.0	2.0	1.0

[0064] 表1

[0065] $$T_k=\frac{\sum_{i=1}^{n}m_i}{\sum_{s=1}^{n}m_s} \quad (5)$$

[0066] 其中，m_i是第i个标签的权值，T_k表示第k个词的平均累加权值，$\sum_{i=1}^{n}m_i$表示第k个词在所在标签的累加权重，$\sum_{s=1}^{n}m_s$表示在整个页面包含的上述标签的权值总和。

[0067] $$sim(g)=\frac{\alpha*\beta}{|\alpha|*|\beta|}=\frac{\sum_{i=1}^{n}g_i*v_i}{\sqrt{(\sum_{i=1}^{n}g_i)^2\ (\sum_{i=1}^{n}V_i)^2}}\times T_k \quad (6)$$

[0068] 式(6)中，r表示主题特征向量的维度，g_i和v_i分别表示特征项t_i在评价页面和主题向量中的权值。待采集页面g的主题相关度sim(g)是一个连续的值，值域是[0,1]。sim(g)不直接认定页面是否与主题相关，而是计算出主题相关的概率。它的值越大表示待抓取页面属于主题页面的概率也就越高。

[0069] 步骤301-2至步骤301-3，具体如下：给定n个相似度的点$\{sim_1, sim_2, \ldots sim_n\}$，找到K 个聚类中心$\{a_1, a_2, \ldots, a_n\}$，使得每个数据点与它最近的聚类中心的距离平方和最小。并将这个距离平方和称为目标函数，记为W_n，其数学表达式为：

[0070] $$W_n = \sum_{i=1}^{n} \min_{1 \le j \le k} | sim_i - a_j |^2 \qquad (7)$$

[0071] (3)计算待抓取页面属于主题类的概率

[0072] 步骤302，根据该待处理网页正文所对应分词集中的各个分词，采用朴素贝叶斯算法计算获得该待处理网页与指定主题关键字的相关概率，具体如下：

[0073] 1)假设D是训练元组和相关联的类标号的集合。每个元组用一个n维属性向量$\{W_1, W_2, W_3 \ldots W_n\}$表示，假设有m个类$C_1, C_2, \ldots, C_m$，并且给定元组$W_x$，即待抓取的网页。朴素贝叶斯分类法预测给定元祖$W_x$是否属于类$C_i$，当且仅当元祖$W_x$属于$C_i$的概率大于元祖$W_x$属于$C_j$的概率。

[0074] $P(C_i|W_x) > P(C_j|W_x)$，$1 \leqslant j \leqslant m, j \neq i$， (8)

[0075] 因此，最大化$P(C_i|X)$的值。其中$P(C_j|X)$最大的类C_i称为最大后验假设.根据贝叶斯定理，得到

[0076] $$P(C_i|W_x) = \frac{P(W_x|C_i)P(C_i)}{P(W_x)} \qquad (9)$$

[0077] 2)由于$P(W_x)$对于所有类为常数，只需要$P(W_x|C_i)P(C_i)$值最大即可.如果这些类的先验概率未知，则通常假定这些类是等概率，即

[0078] $P(C_1) = P(C_2) = \ldots P(C_m)$，并据此对$P(W_x|C_i)$最大化.

[0079] 3)由于计算$P(W_x|C_i)$的开销可能非常大，为降低计算$P(W_x|C_i)$的开销，可以做类的条件独立的朴素假定.给定元组的类标号，假定属性值有条件地相互独立，即在属性之间，不存在依赖关系。这样，

[0080] $$P(W_x|C_i) = \prod_{k=1}^{n} P(W_1|C_i) \times P(W_2|C_i) \times \ldots \times P(W_n|C_i) \qquad (10)$$

[0081] 这样，可以容易地由训练元组的概率$P(W_1|C_i)$，$P(W_2|C_i)$，$\ldots P(W_n|C_i)$。此算法中只需要将结果分为两类，一类是与主题相关的，另一类是与主题的无关的。式子(10)算出的来的结果作为待抓取的网页属于主题类别的阈值。

[0082] 公式(11)中，$Fitness(link_i)$代表第i个网页的连接得分；f_{sim}代表的是第i个网页预测主题相关度，根据公式(6)得出；f_{link}表示的是第i个网页URL的链接分析值，根据公式(12) 得出；HITS算法首先根据查询的关键词确定网络子图$G = (V, E)$，(V为网络子图的结点集，E为边集)，然后通过迭代计算得出每个网页的中心值。通过搜索引擎获得与主题最相关的 K个网页的集合，计算集合中所有页面的中心值和权威值：有向边$<p, q> \in E$，表示页面p 有一条链接指向页面q。然后进行中心值和权威值的计算操作：经过一定次数迭代，直到 A_p和H_p的值收敛。f_{parent}代表的是第i个网页的父页面的相关度，根据公式(6)得出；

$f_{datastruts}$代表的是第i个网页URL的标签权重值，根据公式(5)得出；$f_{relevanturls/totalurls}$代表的是与指定主题关键字相关的网页数量与网页总数量的比值；λ是自适应调整的动态值。

[0083] 步骤8中，针对所有初级待处理网页进行交叉变异操作，获得各个中级待处理网页，其中，遗传算法中的交叉概率的控制对算法的性能有重要的影响。自适应函数比值的变化趋势的判断是通过最近前10次比值的平均值，每进行一次迭代就计算一次平均值。如果完成一次统计，计算得到的前10次的平均值与上一次相比有明显的振幅，系统会根据振幅的正负调整交叉和变异的概率，振幅的区间在[-0.2,+0.2]之间。如果交叉概率过大，就会引入更多新的个体，这种情况容易让系统采集与主题无关的网页，准确率就会降低。而交叉概率过小，又可能使算法早熟收敛，陷入局部极值点。系统不容易抓取新的与主题相关的网页。为了使算法搜索达到全局最优，我们应该在算法运行过程中，依据进化的状态来动态调整交叉概率。采用以下策略：

[0084] $$p_c=\begin{cases} m_1 & ,f_c\le f_{avg} \\ m_2(f_{max}-f_c)/(f_{max}-f_{avg}), & f_c>f_{avg} \end{cases} \quad (13)$$

[0085] 式(13)中p_c表示交叉率，f_{max}表示种群最大适应值，f_{avg}表示种群的平均适应值。f_c表示在要交叉的两个个体中较大的适应度值。判断遗传算法是否收敛到最优值的方法就是看群体平均适应度值与群体最大适应度值之差。另一方面，对于当前群体，适应度值大于平均值的应该适当保留，防止过大的交叉概率破坏最优值附近的解，相反，适应度值小于平均值的，应该尽量交叉，获取最优解。

[0086] 同样，变异概率也可以依据当前进化的状态进行设定。采用以下策略：

[0087] $$p_m=\begin{cases} m_3 & ,f_m\le f_{avg} \\ m_4(f_{max}-f_m)/(f_{max}-f_{avg}), & f_m>f_{avg} \end{cases} \quad (14)$$

[0088] 式(14)中p_m为变异概率，f_m为进行变异操作的个体的适应度值。参照自适应交叉概率来分析当群体是否收敛到局部最优还是全局最优。如果当前个体的适应度函数值大于平均适应度函数值，对该模型适当保护，防止该模型被破坏。相反，我们则利用那些适应度值低于平均值的个体进行完全变异。

[0089] 上述技术方案所设计基于关键字的定向网页采集方法，相较于传统方法，传统基于单机的定向主题采集策略需要消耗大量的时间和带宽，本发明的基于分布式的主题采集策略，利用分布式的方式让多台机器同时对网页进行采集，通过多个节点的并行的定向抓取网络数据，有效提高了数据采集到速率，缩短的采集的时间；而且采集数据准确性是判断采集系统优劣的重要指标之一，本发明的自适应主题采集算法所采集的数据具有较高的数据准确性，能够在降低系统开销的前提下，较为准确地采集与主题相关的网页；并且相较于单纯地人为设定阈值更加合理，能够根据实际的定向采集的与主题网页变化情况自适应调整阈值。以历史采集数据为参考，动态地制定合适的阈值，及时调整系统采集模型，从而实现又好又快的抓取。并且能够在一定程度上提高全局搜索性，避免了采集网页陷入局部最优的状态，通过自适应算法提高系统的整体采集准确率以便于合理地衡量新采集数据的变化程度。

[0090] 上面结合附图对本发明的实施方式作了详细说明，但是本发明并不限于上述实施方式，在本领域普通技术人员所具备的知识范围内，还可以在不脱离本发明宗旨的前提下

做出各种变化。

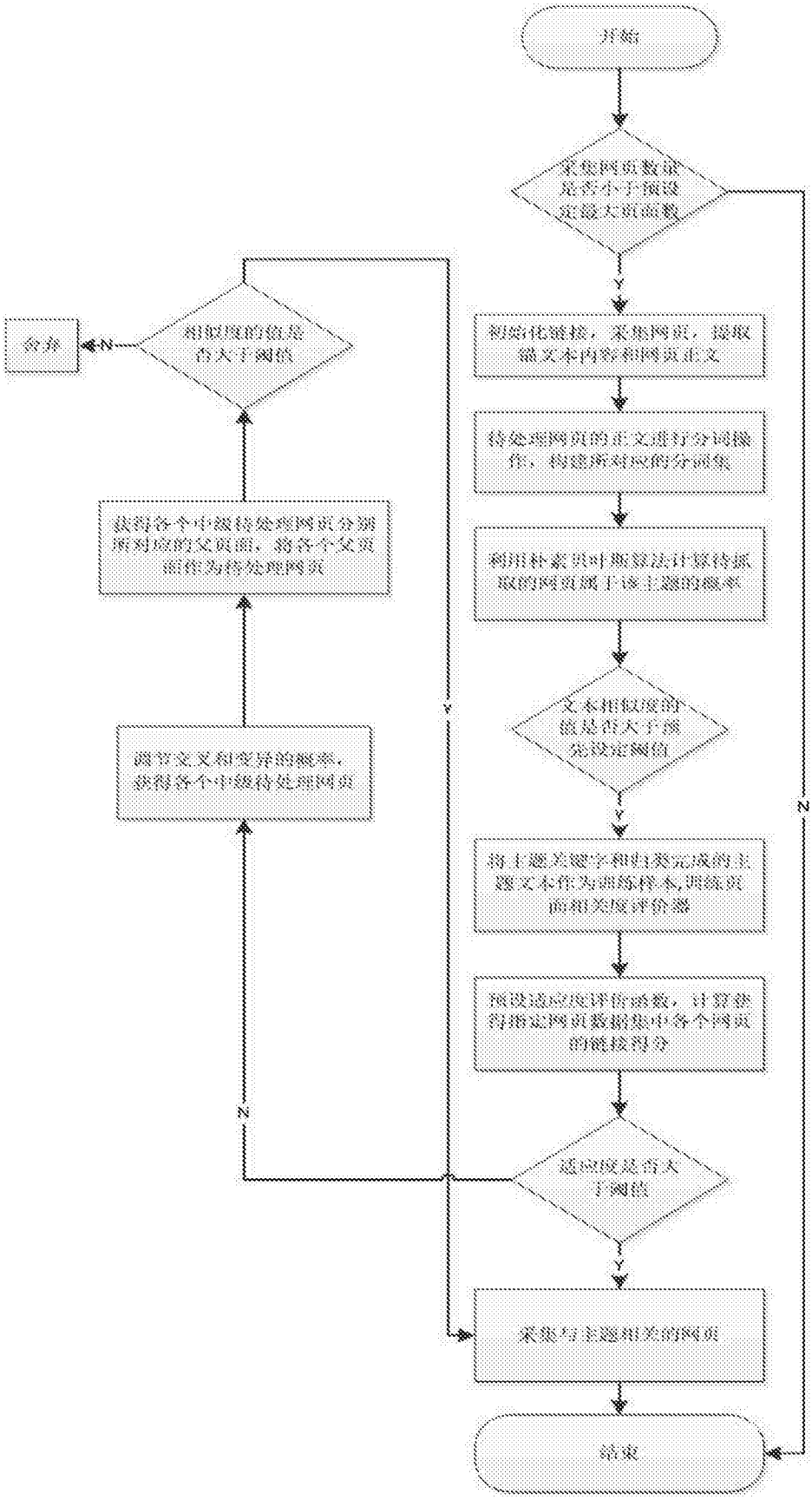

开始
采集网页数量是否小于预设定最大页面数
Y
初始化链接，采集网页，提取锚文本内容和网页正文
待处理网页的正文进行分词操作，构建所对应的分词集
利用林素贝叶斯算法计算待抓取的网页属于该主题的概率
文本相似度的值是否大于预先设定阈值
Y
将主题关键字和归类完成的主题文本作为训练样本，网络页面相关度评价器
预设适应度评价函数，计算获得指定网页数据集中各个网页的链接得分
适应度是否大于阈值
Y
采集与主题相关的网页
结束
N
Y
相似度的值是否大于阈值
N
舍弃
获得各个中级待处理网页分别所对应的父页面，将各个父页面作为待处理网页
调节交叉和变异的概率，获得各个中级待处理网页
N

图1

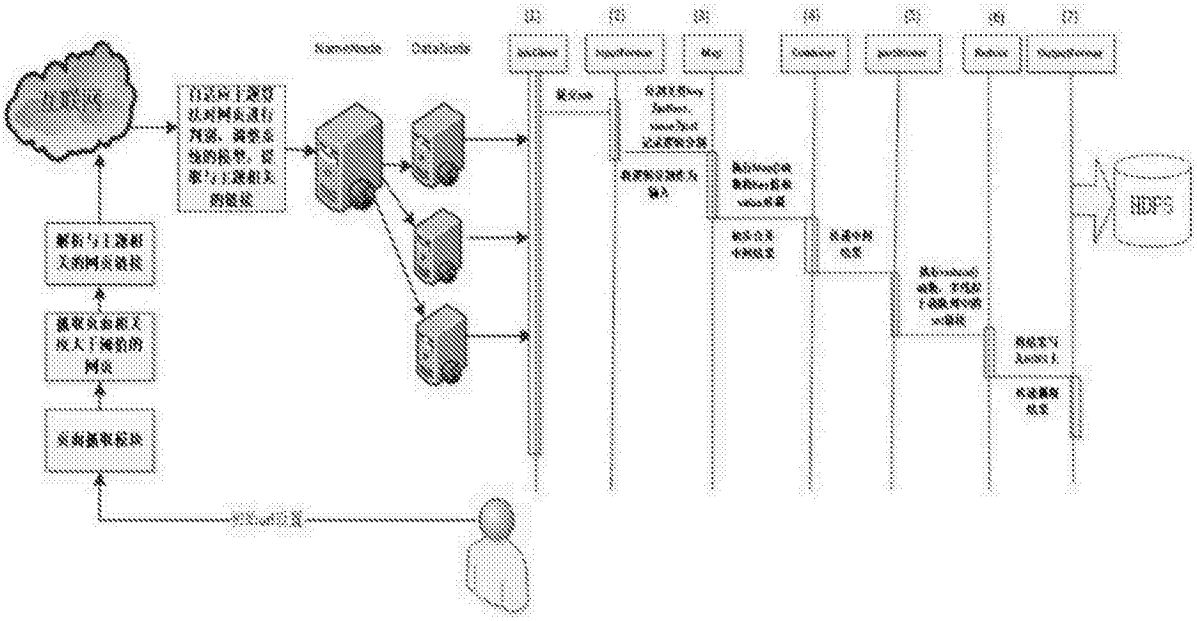


图2