(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 104166683 A
(43)申请公布日 2014.11.26

(21)申请号 201410347539.4

(22)申请日 2014.07.21

(71)申请人 安徽华贞信息科技有限公司
地址 230000 安徽省合肥市高新区黄山路602号国家大学科技园A502

(72)发明人 贾岩

(74)专利代理机构 合肥市长远专利代理事务所（普通合伙） 34119
代理人 程笃庆 黄乐瑜

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

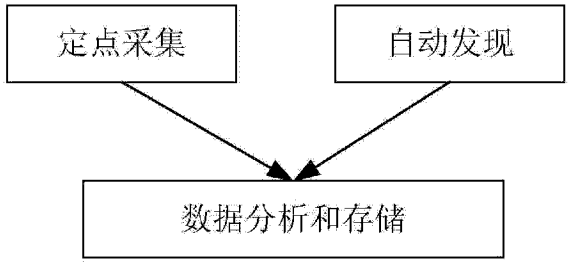
权利要求书1页 说明书3页 附图2页

(54)发明名称

一种数据挖掘方法

(57)摘要

本发明提出了一种数据挖掘方法，解决了网络信息重复程度高，冗余信息多的问题，数据挖掘速率高，查全率好，效果较为理想，其分为，定点采集：预制行业内网站作为数据源，并对每一个数据源设置可信度权值；针对数据源设置数据采集模式定期或不定期向数据源挖掘数据；自动发现：设置网络探针，自动发现相似度高的网站作为采集点网站；将采集点网站添加到采集点网站库，并对每一个采集点网站设置可信度权值；针对采集点网站设置数据提取模式定期或不定期向数据源挖掘数据；数据分析和存储：对挖掘到的数据进行统一编码，去除重复信息，筛选数据；对筛选后的数据进行聚类分析，计算同一话题的信息量，并标注话题关注度权重；存储数据，并建立索引。



CN 104166683 A

1. 一种数据挖掘方法，其特征在于，通过定点采集和自动发现两种方式进行数据挖掘，并对挖掘到的数据进行统一的数据分析和存储；

定点采集包括：

预制行业内网站作为数据源，并对每一个数据源设置可信度权值；

针对数据源设置数据采集模式定期或不定期向数据源挖掘数据；

自动发现包括：

设置网络探针，自动发现相似度高的网站作为采集点网站；

将采集点网站添加到采集点网站库，并对每一个采集点网站设置可信度权值；

针对采集点网站设置数据提取模式定期或不定期向数据源挖掘数据；

数据分析和存储包括：

对挖掘到的数据进行统一编码，去除重复信息，筛选数据；

对筛选后的数据进行聚类分析，计算同一话题的信息量，并标注话题关注度权重；

存储数据，并建立索引。

2. 如权利要求 1 所述的数据挖掘方法，其特征在于，行业内网站包括行业内知名网站链接、论坛、博客。

3. 如权利要求 1 或 2 所述的数据挖掘方法，其特征在于，数据源可信度权值由人工设置。

4. 如权利要求 1 或 2 所述的数据挖掘方法，其特征在于，采集点网站可信度权值人工设置。

5. 如权利要求 1 或 2 所述的数据挖掘方法，其特征在于，采集点网站可信度权值根据网站排名或评分自动设置。

一种数据挖掘方法

技术领域

[0001] 本发明涉及数据挖掘技术领域，尤其涉及一种数据挖掘方法。

背景技术

[0002] 当今社会已经进入信息高速传播的时代，这为人们带来方便的同时，也出现了越来越多的问题，例如，现有搜索引擎搜索结果重复性太高、不符合期望的冗余信息多、搜索时间长、效率低等。

[0003] 由于目前互联网上信息转载率很高，百度、google 等搜索引擎为了搜索的查全率，导致通用搜索耗时长，搜索结果重复度非常高，不利于用户快速发现有价值的内容。另外，一些行业搜索引擎，只针对行业网站，提高了搜索效率，但时查全率低，容易造成遗漏。

[0004] 现在的商业竞争很大程度上决定与企业对最新信息的掌握程度，换言之企业对行业信息的更新与分析决定了企业的潜力，但是企业信息化方面基础千差万别，而且资源都相对有限，尤其是中小企业往往无力承担独立的信息搜索消耗，另一方面，企业定制的搜索引擎往往只搜索行业网站，不对对整个互联网编录，容易造成信息遗漏。

发明内容

[0005] 基于背景技术存在的问题，本发明提出了一种数据挖掘方法，解决了网络信息重复程度高，冗余信息多的问题，数据挖掘速率高，查全率好，效果较为理想。

[0006] 本发明提出的一种数据挖掘方法，通过定点采集和自动发现两种方式进行数据挖掘，并对挖掘到的数据进行统一的数据分析和存储；

[0007] 定点采集包括：

[0008] 预制行业内网站作为数据源，并对每一个数据源设置可信度权值；

[0009] 针对数据源设置数据采集模式定期或不定期向数据源挖掘数据；

[0010] 自动发现包括：

[0011] 设置网络探针，自动发现相似度高的网站作为采集点网站；

[0012] 将采集点网站添加到采集点网站库，并对每一个采集点网站设置可信度权值；

[0013] 针对采集点网站设置数据提取模式定期或不定期向数据源挖掘数据；

[0014] 数据分析和存储包括：

[0015] 对挖掘到的数据进行统一编码，去除重复信息，筛选数据；

[0016] 对筛选后的数据进行聚类分析，计算同一话题的信息量，并标注话题关注度权重；

[0017] 存储数据，并建立索引。

[0018] 优选地，行业内网站包括行业内知名网站链接、论坛、博客。

[0019] 优选地，数据源可信度权值由人工设置。

[0020] 优选地，采集点网站可信度权值人工设置。

[0021] 优选地，采集点网站可信度权值根据网站排名或评分自动设置。

[0022] 本发明即实现了针对行业内网站的重点关注，又兼顾了对整个互联网数据信息的兼顾，前者减少了数据搜索时间，提高了搜索效率，后者提高了搜索结果的查全率，本发明通过二者兼顾的方式，对搜索效率和查全率实现了一个比较理想的平衡。本发明中通过数据统一分析，有效的解决了信息重复的问题，去除冗余信息，减少数据所占空间，同时提高后续处理效率。本发明对数据进行聚类分析并建立索引，可提高数据库的利用效率。

附图说明

[0023] 图 1 为本发明提出的一种数据挖掘方法的流程图；

[0024] 图 2 为定点采集挖掘数据的流程图；

[0025] 图 3 为自动发现挖掘数据的流程图；

[0026] 图 4 为数据分析与存储流程图。

具体实施方式

[0027] 参照图 1，本发明提出的一种数据挖掘方法，通过定点采集和自动发现两种方式进行数据挖掘，并对挖掘到的数据进行统一的数据分析和存储。行业内网站包括行业内知名网站链接、论坛、博客等，定点采集可重点关注这些重要的网站，即关注了行业动态，由缩小了查找网站的时间。自动发现是对定点采集的补充，通过对其他非知名网站的搜索，补充数据，避免目标数据的遗漏。数据统一分析可有效去除重复信息，解决了网络数据转载频繁，信息重复的问题，同时。

[0028] 参照图 2，定点采集包括以下步骤：

[0029] 预制行业内网站作为数据源，并对每一个数据源设置可信度权值；

[0030] 针对数据源设置数据采集模式定期或不定期向数据源挖掘数据。

[0031] 数据源预制，即节约了网站搜索时间，提高数据采集效率，又提高了行业针对性，使得采集数据的方向更加符合用户预期。数据源可信度权值由人工设置，可作为数据采集的参考。

[0032] 参照图 3，自动发现包括以下步骤：

[0033] 设置网络探针，自动发现相似度高的网站作为采集点网站；

[0034] 将采集点网站添加到采集点网站库，并对每一个采集点网站设置可信度权值；

[0035] 针对采集点网站设置数据提取模式定期或不定期向数据源挖掘数据。

[0036] 网络探针的设置以数据源为参考，如此可限定探针发现网站的方向，缩小采集点网站的范围，减小数据挖掘范围，提高速率并减少存储空间，同时，数据源为参考也可以提高采集点网站与行业信息的相关度，减少冗余信息。

[0037] 本实施方式中，采集点网站可信度权值根据网站排名或评分自动设置，考虑到网络的发达，各种网站繁杂纷乱，自动设置可减少人力需求并提高工作效率。具体实施时，采集点网站可信度权值也可人工设置，该种方式更加符合用户期望，数据采集精度更高。

[0038] 参照图 4 数据分析和存储包括以下步骤：

[0039] 对挖掘到的数据进行统一编码，去除重复信息，筛选数据；

[0040] 对筛选后的数据进行聚类分析，计算同一话题的信息量，并标注话题关注度权重；

[0041] 存储数据，并建立索引。

[0042] 本实施方式中，有效的解决了信息重复的问题，减少数据所占空间，同时提高后续处理效率。对数据进行聚类分析并建立索引，可提高数据库的检索效率，提高数据利用率。话题关注度的计算与标注，明确提醒用户关注重要信息。

[0043] 以上所述，仅为本发明较佳的具体实施方式，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，根据本发明的技术方案及其发明构思加以等同替换或改变，都应涵盖在本发明的保护范围之内。

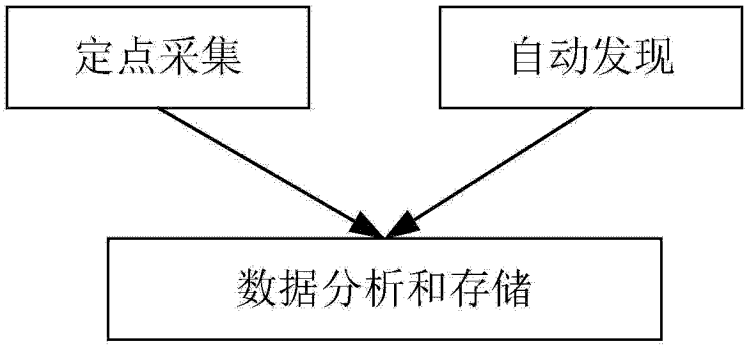

定点采集
自动发现
数据分析和存储

图 1

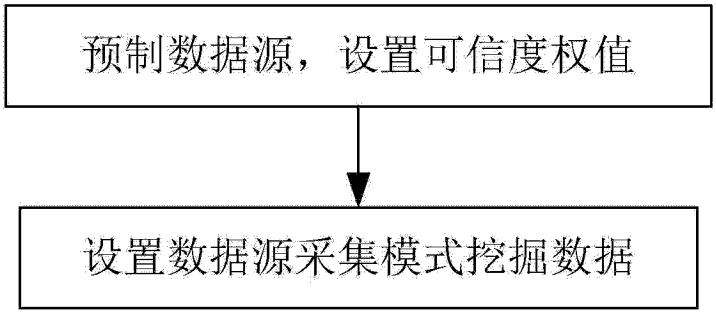

预制数据源，设置可信度权值
设置数据源采集模式挖掘数据

图 2

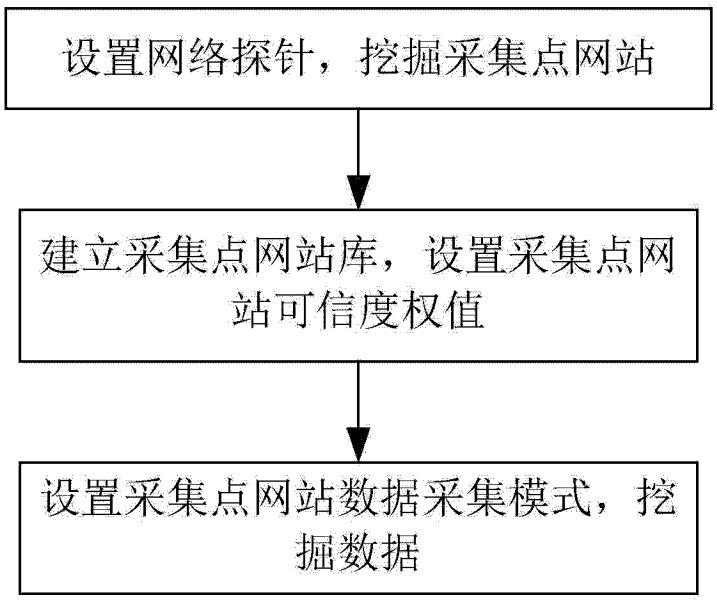

设置网络探针，挖掘采集点网站
建立采集点网站库，设置采集点网站可信度权值
设置采集点网站数据采集模式，挖掘数据

图 3

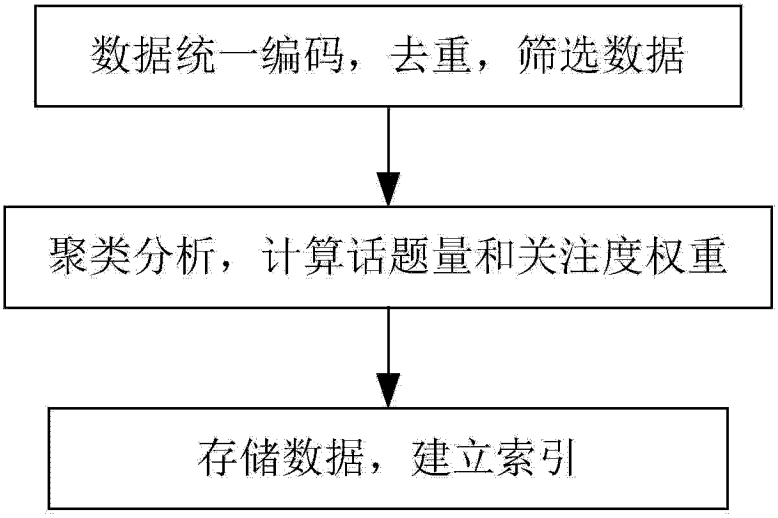


图 4