(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102831253 B
(45)授权公告日 2015.01.21

(21)申请号 201210362934.0

(22)申请日 2012.09.25

(73)专利权人 北京科东电力控制系统有限责任公司
地址 100192 北京市海淀区清河小营东路15号
专利权人 华中电网有限公司

(72)发明人 何蕾 李勇 曹宇 喻宏元 苏迤 庞传军 聂春元 杨笑宇 徐家慧 武毅 林海峰 方伟

(74)专利代理机构 北京金智普华知识产权代理有限公司 11401
代理人 皋吉甫

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

(56)对比文件
CN 101561815 A,2009.10.21, 全文.
CN 101853288 A,2010.10.06, 全文.
CN 101789006 A,2010.07.28, 全文.
CN 102054009 A,2011.05.11, 全文.
WO 2007079303 A3,2007.08.23, 全文.
林乐然等.基于云计算的分布式企业搜索引擎研究.《电脑知识与技术》.2009, 第5卷(第33期), 第9430页.
唐华姣等.基于Lucene的分布式并行索引.《计算机技术与发展》.2011, 第2卷(第2期), 第125页左栏,图1.

审查员 张晓芳

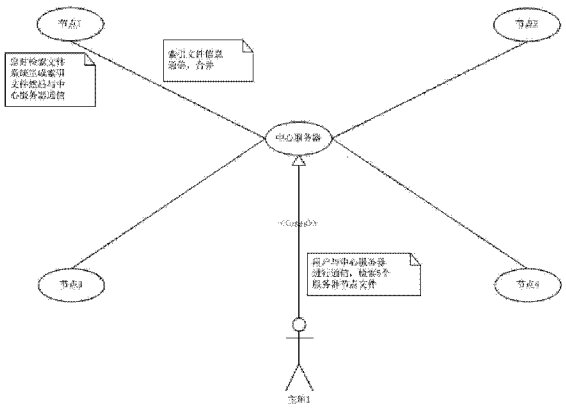
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种分布式全文检索系统

(57)摘要

本发明属于数据处理领域,特别是涉及一种分布式全文检索系统。所述系统由设置在各网络节点上的全文检索服务器组成,包括分布式节点服务器及系统中心服务器;所述分布式节点服务器包括文件信息检索模块及服务器间通讯模块;所述文件信息检索模块对本节点服务器上文件定时进行全文信息的检索,按照定义好的词库进行切词,提取关键词信息并建立索引文件;所述服务器间通讯模块实现分布式节点服务器与系统中心服务器与之间的信息交换,所述系统中心服务器对各节点服务器传送的索引文件进行合并,向用户提供查询接口,将符合的文件作为查询结果展示给用户。



CN 102831253 B

1. 一种分布式全文检索系统，所述系统由设置在各网络节点上的全文检索服务器组成，包括分布式节点服务器及系统中心服务器；其特征在于：

所述分布式节点服务器包括文件信息检索模块及服务器间通讯模块；

所述文件信息检索模块对本节点服务器上文件定时进行全文信息的检索，按照定义好的词库进行切词，提取关键词信息并建立索引文件；所述索引文件包括文件名称、关键词条、文件分类、所在服务器信息、文件大小、文件作者相关信息，所述服务器信息包括服务器的 IP 地址；

所述服务器间通讯模块实现分布式节点服务器与系统中心服务器之间的信息交换，将本节点的索引文件发送到系统中心服务器；

所述系统中心服务器对各节点服务器传送的索引文件进行合并，生成新的索引文件并更新已有索引信息，增加新的文件信息；向用户提供查询接口，响应用户查询文件的请求、分析用户请求，将查询关键词在新的索引文件中进行检索比对，将符合的文件作为查询结果展示给用户；

所述分布式节点服务器还包括：词库管理模块，所述词库管理模块在遍历文件全文内容时根据已有词库进行切词划分，将文件内容切成不同的关键词，然后统计关键词出现的频度和关键词的分类，一同写入到索引文件中；所述词库管理模块按照电网相关技术知识进行统计划分，包括电网文件类、技术论文类、电网设备类、新闻类；对普通的助词、语气词或普通描述性的词进行过滤；所述分布式节点服务器上安装词库管理客户端，所述客户端对在文件中出现频度较高的词，通过用户手动维护添加到索引文件中更新词库。

2. 根据权利要求 1 所述的一种分布式全文检索系统，其特征在于，所述系统中心服务器进一步包括：

各个分布式节点服务器状态查询模块及系统重启服务模块；并具有文件在线浏览与下载模块，即系统中心服务器接收下载文件请求，并根据该文件在索引文件中的描述，将请求转发给相应的节点服务器，将读取文件的字节流返给用户实现下载。

一种分布式全文检索系统

技术领域

[0001] 本发明属于数据处理领域，特别是涉及一种分布式全文检索系统的。

背景技术

[0002] 文档检索系统主要实现对调度管理应用中的各类文档、资料和知识库的索引提取及资料搜索功能。

[0003] a) 编制索引：对 doc、txt、pdf 等常用文档文件的文字信息进行文字索引提取。

[0004] b) 资料搜索：根据文字索引模糊搜索定位文档、资料。

[0005] 分布式查询主要实现调度机构之间、调度管理类应用中各模块标准化的数据库信息分布式查询。基于平台的远程服务代理和数据公共服务实现远程数据查询。分布式查询应包含但不限于以下功能：

[0006] c) 跨调度机构的数据查询；

[0007] d) 按照数据分类进行查询；

[0008] e) 数据展示功能。

[0009] 由于电力行业相关文件，电子信息材料，新闻应用比较多，很多系统都有自己的管理文档的功能，但是面对大量的信息资源，很难定位想要找的文件在那个系统中，存储在哪个服务器上，如何方便快捷、准确地从各个分布式的服务器上上获取所需文件信息，成为至关重要的问题。现有技术中的文档检索系统存在着检索速度慢、占用系统资源过多等的缺陷。

发明内容

[0010] 本发明的目的，是提供一种分布式全文检索系统，从而实现提升检索速度，达到优化目的。

[0011] 本发明的具体技术方案如下：1、一种分布式全文检索系统，所述系统由设置在各网络节点上的全文检索服务器组成，包括分布式节点服务器及系统中心服务器；

[0012] 所述分布式节点服务器包括文件信息检索模块及服务器间通讯模块；

[0013] 所述文件信息检索模块对本节点服务器上文件定时进行全文信息的检索，按照定义好的词库进行切词，提取关键词信息并建立索引文件；

[0014] 所述服务器间通讯模块实现分布式节点服务器与系统中心服务器与之间的信息交换，将本节点的索引文件发送到系统中心服务器；

[0015] 所述系统中心服务器对各节点服务器传送的索引文件进行合并，生成新的索引文件并更新已有索引信息，增加新的文件信息；向用户提供查询接口，响应用户查询文件的请求、分析用户请求，将查询关键词在新的索引文件中进行检索比对，将符合的文件作为查询结果展示给用户。

[0016] 进一步的，所述分布式节点服务器上生成的索引文件包括文件名称、关键词条、文件分类、所在服务器的 IP 地址、服务器信息、文件大小、文件作者等相关信息。

[0017] 进一步的，所述系统中心服务器进一步包括：各个分布式节点服务器状态查询模块及系统重启服务模块；并具有文件在线浏览与下载模块，即系统中心服务器接收下载文件请求，并根据该文件在索引文件中的描述，将请求转发给相应的节点服务器，将读取文件的字节流返给用户实现下载。

[0018] 进一步的，所述分布式节点服务器还包括：词库管理模块，所述词库管理模块在遍历文件全文内容时根据已有词库进行切词划分，将文件内容切成不同的关键词，然后统计关键词出现的频度和关键词的分类，一同写入到索引文件中。

[0019] 进一步的，所述词库管理模块按照电网相关技术知识进行统计划分，包括电网文件类、技术论文类、电网设备类、新闻类等；对普通的助词、语气词或普通描述性的词进行过滤。

[0020] 进一步的，所述分布式节点服务器上安装词库管理客户端，所述客户端对在文件中出现频度较高的词，通过用户手动维护的关键词等添加到索引文件中更新词库。

[0021] 本发明的有益效果是：

[0022] （1）在查询效率上，由于使用了依据电力行业知识的词库管理，在生成索引文件时就会过滤一些不明感的、不关心词汇，减小生成的索引文件，提高检索速度。

[0023] （2）提供了各节点管理的界面，可以维护各节点相关文件配置，索引生成、词库管理以及与服务器通信等功能，加强了分布式系统的稳定性。

[0024] （3）索引文件格式独立于应用平台，定义了一套以 8 位字节为基础的索引文件格式，使得兼容系统或者不同平台的应用能够共享建立的索引文件。

[0025] （4）在传统全文检索引擎的倒排索引的基础上，实现了分块索引，能够针对新的文件建立小文件索引，提升索引速度。然后通过与原有索引的合并，达到优化的目的。

[0026] （5）实现了一套强大的查询引擎，默认实现了布尔操作、模糊查询、分组查询等等。

附图说明

[0027] 图 1 是本发明的系统结构框图。

具体实施方式

[0028] 下面具体阐述本发明的技术方案。

[0029] 本发明针对于常规分布式策略，如果在一个机器上没有找到匹配的文件，则将用户请求转发到其他机器上继续检索索引文件。这样每次请求都会遍历所有机器的索引文件，效率以及负载较大，该系统采用将各节点索引文件统一到一个中心服务器机器上，减少转发请求的时间，同时只在中心服务器上进行检索，减轻其他节点机器的负载如图 1 所示承担一种分布式全文检索系统，由设置在各网络节点上的全文检索服务器组成，按照功能划分又分为分布式节点服务器及系统中心服务器，图 1 中包括一系统中心服务器及 4 个分布式节点服务器。其中，每个节点服务器包括文件信息检索模块及服务器间通讯模块；文件信息检索模块对本节点服务器上文件定时进行全文信息的检索，并且频率可设置，按照定义好的词库进行切词，提取关键词信息并建立索引文件。生成的索引文件包括文件名称、关键词条、文件分类、所在服务器的 IP 地址、服务器信息、文件大小、文件作者等相关信息。服务器间通讯模块则实现分布式节点服务器与系统中心服务器与之间的信息交换，包括将本

节点的索引文件发送到系统中心服务器，或者相应来自系统中心服务器的用户请求等。

[0030] 系统中心服务器对各节点服务器传送的索引文件进行合并，生成新的索引文件，并且在此基础上不断更新已有索引信息，增加新的文件信息。向用户提供查询接口，响应用户查询文件的请求、分析用户请求，将查询关键词在新的索引文件中进行检索比对，将符合的文件作为查询结果展示给用户。

[0031] 进一步的，本发明的系统中心服务器进一步包括：包括各个节点索引文件信息，管理节点与中心服务器通信，配置各个节点文件索引信息等，并提供重启服务等功能模块；并具有文件在线浏览与下载模块，即系统中心服务器接收下载文件请求，并根据该文件在索引文件中的描述，将请求转发给相应的节点服务器，将读取文件的字节流返给用户实现下载。

[0032] 本发明为了进一步提供查询的效率，在分布式节点服务器还包括词库管理模块，该词库管理模块在遍历文件全文内容时根据已有词库进行切词划分，将文件内容切成不同的关键词，然后统计关键词出现的频度和关键词的分类，一同写入到索引文件中。词库管理模块按照电网相关技术知识进行统计划分，包括电网文件类、技术论文类、电网设备类、新闻类等；对普通的助词、语气词或普通描述性的词进行过滤。

[0033] 在本系统的分布式节点服务器上还安装词库管理客户端，客户端对在文件中出现频度较高的词，通过用户手动维护的关键词等添加到索引文件中更新词库。更进一步提高了查询的效率，更具有针对性。

[0034] 该分布式全文检索系统的使用，大大提高了用户查找所需文件的效率。而且，针对现有的搜索引擎在搜索效率、信息维护、分布式节点管理、负载压力等方面存在的问题有了提高与优化。

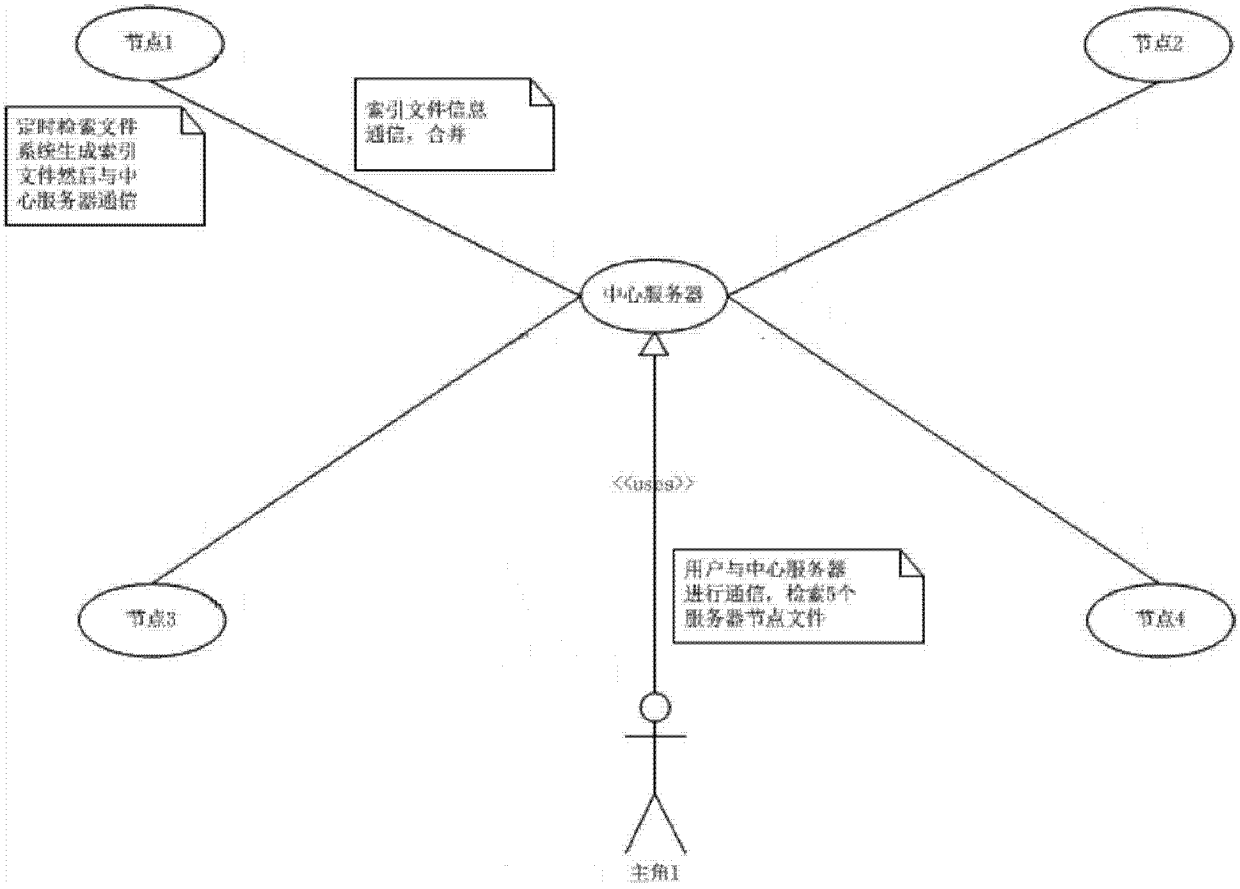

节点1
节点2
定时检索文件
系统生成索引
文件然后与中
心服务器通信
索引文件信息
通信，合并
中心服务器
<<uses>>
用户与中心服务器
进行通信，检索5个
服务器节点文件
节点3
节点4
主角1

图 1