



(12) 发明专利申请

(10) 申请公布号 CN 102810089 A

(43) 申请公布日 2012. 12. 05

(21) 申请号 201110142354. 6

(22) 申请日 2011. 05. 30

(71) 申请人 盛乐信息技术(上海)有限公司  
地址 201203 上海市浦东新区郭守敬路 356 号

(72) 发明人 范路

(74) 专利代理机构 上海浦一知识产权代理有限公司 31211  
代理人 高月红

(51) Int. Cl.  
G06F 17/30(2006. 01)

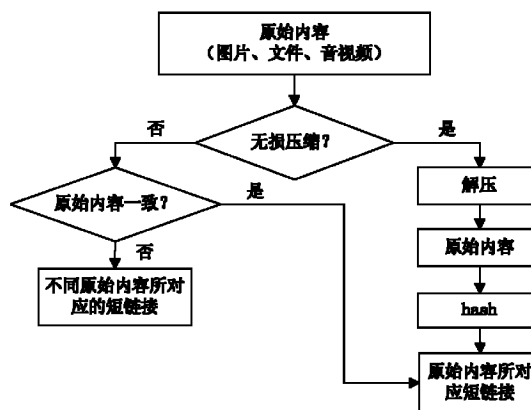
权利要求书 1 页 说明书 4 页 附图 1 页

(54) 发明名称

基于内容的短链接系统及实现方法

(57) 摘要

本发明公开了一种基于内容的短链接系统及实现方法,该系统包括:数据生成模块、短链接生成模块、链接存储模块;其实现方法,包括:1) 将需要缩短的地址的原始内容,生成 hash 数据;2) 将 hash 数据经短链接生成模块,形成原始内容所对应的短链接;3) 将原始内容的短链接和原始内容的地址链接存储到链接存储模块;4) 当用户拿着短链接进行解析时,短链接网址会根据链接存储模块中的 key-value 进行检索,并将原来的网址链接返回回来,进行地址重定向。本发明实现了多个链接对应的实际内容相同时的负载均衡及实时确定短链接的有效性,方便用户使用。



1. 一种基于内容的短链接系统,其特征在于:包括:  
数据生成模块,用于将原始内容经 hash 运算,生成 hash 数据;  
短链接生成模块,用于将 hash 数据经多进制数字的编码进行压缩,形成原始内容所对应的短链接;  
链接存储模块,用于对原始内容的短链接和原始内容的地址链接进行存储。
2. 如权利要求 1 所述的系统,其特征在于:所述原始内容包括:图片、文件、音频、视频的内容。
3. 如权利要求 1 所述的系统,其特征在于:所述多进制数字,包括:一套允许在 url 地址上出现的字符形成的多进制数字。
4. 如权利要求 1 所述的系统,其特征在于:所述链接存储模块中,以 key-value 数据库的方式进行存储。
5. 如权利要求 1 所述的系统,其特征在于:还包括:检查模块,用于定期检查短链接库中所存放的原始链接所指向的内容,在内容发生变化之后,自动失效地址库中的短链接到该原始链接的指向。
6. 如权利要求 1 所述的基于内容的短链接系统的实现方法,包括步骤:
  - (1) 将需要缩短的地址的原始内容,经数据生成模块,生成 hash 数据;
  - (2) 将 hash 数据经短链接生成模块,形成原始内容所对应的短链接;
  - (3) 将原始内容的短链接和原始内容的地址链接存储到链接存储模块;
  - (4) 当用户拿着短链接进行解析时,短链接网址会根据链接存储模块中的 key-value 进行检索,并将原来的网址链接返还回来,进行地址重定向。
7. 如权利要求 6 所述的实现方法,其特征在于:所述短链接中,根据原始内容使用的是否为无损压缩源,进行以下的短链接生成处理:
  - (1) 如果原始内容使用的是无损压缩源,则将压缩内容进行解压,生成原始内容数据,再对这些数据进行 hash 数据处理,再使用多进制数字进行压缩,形成原始内容所对应的短链接;
  - (2) 如果原始内容使用的是有损压缩源,那么如果原始内容一致,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成一致的原始内容所对应的短链接;  
如果原始内容不一致,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成不同的原始内容所对应的短链接。
8. 如权利要求 6 所述的实现方法,其特征在于:还包括步骤:(5) 检查模块定期扫描短链接库中所存放的原始链接所指向的内容,在内容发生变化之后,自动失效地址库中短链接到该链接的指向。

## 基于内容的短链接系统及实现方法

### 技术领域

[0001] 本发明涉及一种短链接系统及实现方法,特别是涉及一种基于内容的短链接系统及实现方法。

### 背景技术

[0002] 微博 (MicroBlog) 是一个基于用户关系的信息分享、传播以及获取平台,用户可以通过 web(互联网)、wap(Wireless Application Protocol,无线应用协议)以及各种客户端组件个人社区,以 140 字左右的文字更新信息,并实现即时分享。由于微博具有便捷性、原创性等特点,应用非常广泛。

[0003] 随着微博的流行,由于其内容有着字数的限制,所以大家纷纷采用短链接来解决引用链接过长,占掉有效字数的问题。

[0004] 现在的短链接,都是基于网址的。为了表达越来越丰富的内容,网址链接越来越长。现有的短链接服务,会把需要缩短的网址,进行 hash 离散运算,得到一串数字之后,再使用多进制数字,来表达 hash 之后的唯一结果。然后使用 key-value 数据库对短链接和原网址链接进行存储。但是,现在的短链接通常是使用链接为唯一标示的,也就是说一个短链接对应一个链接。如果有多个链接对应的实际内容是相同的,特别是一些多媒体信息的时候,这样就不利于内容访问的容错和负载均衡。

### 发明内容

[0005] 本发明要解决的技术问题是提供一种基于内容的短链接系统及实现方法,以实现多个链接对应的实际内容相同时的负载均衡。

[0006] 为解决上述技术问题,本发明的基于内容的短链接系统,包括:

[0007] 数据生成模块,用于将原始内容经 hash 运算,生成 hash(哈希)数据;其中,该原始内容包括:图片、文件、音频、视频等的内容;

[0008] 短链接生成模块,用于将 hash 数据经多进制数字的编码进行压缩,形成原始内容所对应的短链接;

[0009] 链接存储模块,用于对原始内容的短链接和原始内容的地址链接进行存储,如可以 key-value 数据库的方式进行存储。

[0010] 所述多进制数字,包括:一套允许在 url 地址上出现的字符形成的多进制数字。

[0011] 另外,本发明的基于内容的短链接系统,还包括:检查模块,用于定期检查短链接库中所存放的原始链接所指向的内容,在内容发生变化之后,自动失效地址库中的短链接到该原始链接的指向。

[0012] 再者,本发明的基于内容的短链接系统的实现方法,包括步骤:

[0013] (1) 将需要缩短的地址的原始内容,经数据生成模块,生成 hash 数据;

[0014] (2) 将 hash 数据经短链接生成模块,形成原始内容所对应的短链接;

[0015] (3) 将原始内容的短链接和原始内容的地址链接存储到链接存储模块;

[0016] (4) 当用户拿着短链接进行解析时,短链接网址会根据链接存储模块中的 key-value 进行检索,并将原来的网址链接返还回来,进行地址重定向。

[0017] 所述短链接中,根据原始内容使用的是否为无损压缩源,进行以下的短链接生成处理:

[0018] (1) 如果原始内容使用的是无损压缩源,则将压缩内容进行解压,生成原始内容数据,再对这些数据进行 hash 数据处理,再使用多进制数字进行压缩,形成原始内容所对应的短链接;

[0019] (2) 如果原始内容使用的是有损压缩源,那么如果原始内容一致,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成一致的原始内容所对应的短链接;

[0020] 如果原始内容不一致的话,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成不同的原始内容所对应的短链接。

[0021] 上述方法,还可包括步骤:(5) 检查模块定期扫描短链接库中所存放的原始链接所指向的内容,在内容发生变化之后,自动失效地址库中短链接到该链接的指向。

[0022] 本发明通过基于目标地址的实际内容进行 hash 运算,然后根据 hash 的结果,生成短链接。如果目标地址的内容是相同的,那么生成的 hash 结果也应该是相同的,也就可以分配相同的短链接。当有用户需要访问这些短链接的时候,短链接服务器随机或根据负载均衡算法,给出多个地址中的一个,实现了负载均衡的目的。因此,本发明与传统的短链接的差异在于,对于相同的内容,其短链接地址是相同的。

[0023] 另外,本系统还有一个流程,就是地址有效性检查。传统的短链接,是没法对地址的有效性进行检查的,因为他们不记录原地址的内容。而本本发明的系统可定期扫描短链接库中所存放的原始链接所指向的内容,如果内容不一致了,则删除这个原地址,以便实时确定短链接的有效性,方便用户使用。

## 附图说明

[0024] 下面结合附图与具体实施方式对本发明作进一步详细的说明:

[0025] 图 1 是本发明的原始内容所对应的短链接的形成示意图;

[0026] 图 2 是本发明的短链接的解析示意图。

## 具体实施方式

[0027] 本发明的基于内容的短链接系统,包括:

[0028] 数据生成模块,用于将网址上的原始内容经 hash 运算,生成 hash 数据;其中,该原始内容包括:图片、文件、音频、视频等的内容;关于 hash 运算,可按现有的 hash 运算进行,例如,可参考 <http://zh.wikipedia.org/wiki/MD5> 所介绍的那样;

[0029] 短链接生成模块,用于将 hash 数据经多进制数字的编码进行压缩,形成原始内容所对应的短链接;

[0030] 链接存储模块,用于将原始内容的短链接和原始内容的地址链接,存储到 key-value 数据库;

[0031] 检查模块,用于定期检查短链接库中所存放的原始链接所指向的内容,在内容发生变化之后,自动失效地址库中的短链接到该原始链接的指向。

[0032] 其中,关于多进制数字:同样一个数字,二进制的位数最长,因此,二进制的要比四进制的位数多,十进制的会比四进制、八进制的短,十六进制的更短。每一个位数所能够表达的变化越多,同一个数字所需占用的位数就越短。比如 80 进制,每一位能够表达 80 种变化,其长度,肯定要比直接原来 MD5 算法所得到的 16 进制的数字短很多。因此,本实施例中,可以采用以下的一套多进制数字,来缩短 hash 结果的长度:

[0033] 使用所有允许在 url 地址 (Uniform/Universal Resource Locator, 网页地址) 上出现的字符,来形成一套多进制的数字,比如 0 ~ 9, a ~ z 再加上 A ~ Z, 也就是 62 个字符,形成一套 62 进制的编码,再加上允许在 url 中出现的“/ . % ? &”等特殊字符,这个编码的进制可以进一步的增加,长度可以进一步的降低。那么就可以将同一段数字的长度进一步的缩短。

[0034] 比如,80 在二进制中是一个 7 位数,如果是二进制中的 7 个 1 的话,这个数字就是 128,但允许出现在 url 中的字符又没有那么,因此,可以采用 6 位二进制数,也就是 64 种变化。使用 64 的字符来代表这 64 种变化,其中大小写字母再加上十个数字,加在一起,是 62 个,再加上在 url 中比较常见的下划线“\_”和百分号“%”,正好 64 个。如果有些 web 服务器或浏览器有兼容性问题的话,也可以换成“#”或“?”等字符。

[0035] 而对于目前最常用的多进制舍去算法,形成短链接的处理过程,现以 jpg 格式存储的图片为例,说明如下:

[0036] (1) 先将图片的数据读入内存,然后对其进行 hash 转换成为一个 128 位二进制的数字;

[0037] (2) 然后再对这个二进制的数字,进行多位编码,具体如下:

[0038] 1) 将长网址 md5 生成 32 位签名串,分为 4 段,每段 8 个字节;

[0039] 2) 对这四段循环处理,取 8 个字节,将它看成 16 进制串与  $0 \times 3\text{fffffff}$  (30 位 1) 与操作,即超过 30 位的忽略处理;

[0040] 3) 将这 30 位分成 5 段,每 6 位的数字作为字母表的索引取得特定字符,依次进行获得 5 位字符串;

[0041] 4) 总的 md5 串可以获得 4 个 5 位串;取里面的任意一个就可作为这个长 url 的短 url 地址。

[0042] 因此,本实施例中采用的多进制数字算法,相对于现有的最常用的多进制舍去算法,字符的变化更加多。

[0043] 在短链接中,本发明的系统可根据原始内容使用的是否为无损压缩源,进行以下的短链接生成处理(如图 1 所示):

[0044] (1) 如果原始内容使用的是无损压缩源,也就是可以还原出唯一的无压缩内容;将压缩内容进行解压,生成原始内容数据,再对这些数据进行 hash 数据处理,再使用上述的多进制数字进行压缩,形成原始内容所对应的短链接;

[0045] 其中,对于无损压缩的原始内容,即使格式不同,只要解压后的原始数据一致,那么得到的短链接也是一样的;

[0046] (2) 如果原始内容使用的是有损压缩源,则只对原始内容进行比较,如果原始内容一致,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成一致的原始内容所对应的短链接;

[0047] 如果原始内容不一致的话,在将原始内容进行 hash 数据处理,再使用多进制数字进行压缩后,形成不同的原始内容所对应的短链接。

[0048] 对于上述的基于内容的短链接系统,其实现方法,包括步骤:

[0049] (1) 将需要缩短的地址的原始内容,经数据生成模块,生成 hash 数据;

[0050] (2) 将 hash 数据经短链接生成模块,形成原始内容所对应的短链接;

[0051] (3) 将原始内容的短链接和原始内容的地址链接存储到链接存储模块(即 key-value 数据库);

[0052] (4) 当用户拿着短链接进行解析时,短链接网址会根据链接存储模块中的 key-value 进行检索,并将原来的网址链接返还回来,进行地址重定向;

[0053] (5) 检查模块定期的对记录的原地址进行检查,重新得到地址所指向的内容,并将内容进行 hash 运算之后,校验一下是否和储存的样本一致。如果不一致了,则删除这个原地址。再有用户使用这个短链接的时候,那个损坏掉的原地址,就不会被交付给用户了。如果关于一个内容的所有原地址都损坏了,则不再进行解析,直接在用户申请的时候,返回失败信息。

[0054] 本发明中,为了提高短链接的解析速度,采用 key-value 结构来进行数据存储。

[0055] 一个 collection 以 hash 结果为 key,以短链接为 value,另外一个,则以短链接为 key,以原始链接的列表为 value。当用户存入一个短链接的时候,先进行 hash 运算,然后到第一个 collection 中去寻找 key,如果没找到,则同时插入 resource2url 和 url2url(图 2)。如果在 resource2url 中找到了 hash 结果的 key,那么就直接将 resource2url 中的 value 也就是短链接拿出来作为 key 到 url2url 中去取出 value, source-url-list,然后将新的 url 追加到后面去。

[0056] 在用户拿着短链接进行解析的时候,只要到 url2url 中,根据短链接的 key,取出 source-url-list,然后根据分隔符(比如是 '{ }' 或 '<>'),将多个地址解析出来,再随机给出其中一个。

[0057] 当系统进行原始地址复查的时候,则需要遍历 url2url,将原始地址所指向的内容提取出来,进行 hash 之后,与存放在 url2url 中的 resource-hash 进行比较,如果相同,则说明内容有效;如果不同,则说明原始链接无效了。在验证结束之后,将所有通过验证的地址,重新写入 source-url-list。

[0058] 综上所述,采用本发明的系统,当用户进行短链接地址解析的时候,短链接服务器会自动的进行负载均衡。负载均衡,可以采用最简单的随机数算法,当一个短链接被请求的时候,短链接服务器,随机的给出一个原内容的地址,进行跳转。针对同一个内容的不同资源,其被调用的机会是均等的,即实现了负载均衡的目的。另外,通过本发明系统的定期短链接地址的有效性检查,了解原始内容的变化,方便用户使用。

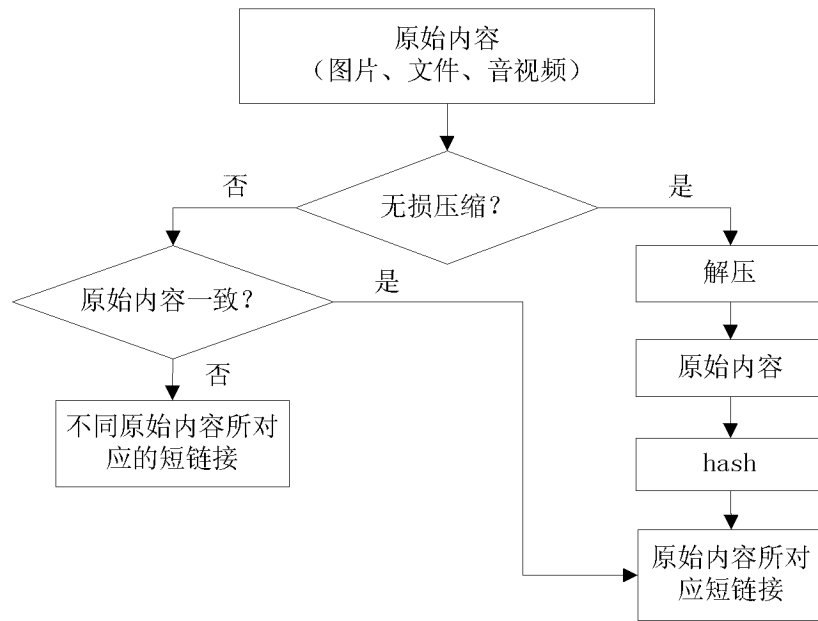


图 1

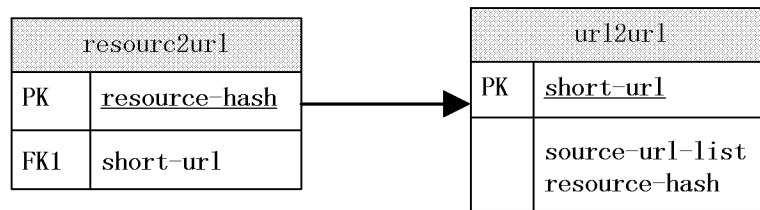


图 2