(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111400563 B
(45)授权公告日 2023.08.01

(21)申请号 202010183402.5

(22)申请日 2020.03.16

(65)同一申请的已公布的文献号
申请公布号 CN 111400563 A

(43)申请公布日 2020.07.10

(73)专利权人 北京搜狗科技发展有限公司
地址 100084 北京市海淀区中关村东路1号院9号楼搜狐网络大厦9层01房间

(72)发明人 孙浩

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
专利代理师 郑傲日

(51)Int.Cl.
G06F 16/903(2019.01)
G06F 40/289(2020.01)

(56)对比文件
CN 110276071 A,2019.09.24
US 2015358264 A1,2015.12.10

审查员 田晶

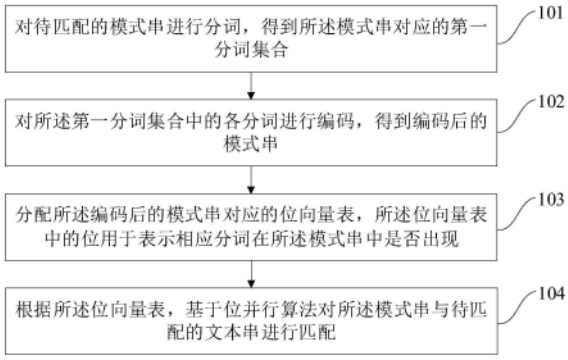
权利要求书4页 说明书16页 附图3页

(54)发明名称

一种模式匹配方法、装置和用于模式匹配的装置

(57)摘要

本发明实施例提供了一种模式匹配方法、装置和用于模式匹配的装置。其中的方法具体包括:对待匹配的模式串进行分词,得到所述模式串对应的第一分词集合;对所述第一分词集合中的各分词进行编码,得到编码后的模式串;分配所述编码后的模式串对应的位向量表,所述位向量表中的位用于表示相应分词在所述模式串中是否出现;根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配。本发明实施例可以提高B表的利用率,以及提高模式匹配的效率。



CN 111400563 B

1.一种模式匹配方法,其特征在于,所述方法包括:

对待匹配的模式串进行分词,得到所述模式串对应的第一分词集合;

对所述第一分词集合中的各分词进行编码,得到编码后的模式串;

在所述编码后的模式串的长度大于预设的分组长度时,按照所述分组长度,对所述编码后的模式串进行划分,得到划分后的各个子串;

在所述各个子串中,确定目标子串;

分配所述编码后的模式串对应的位向量表,所述位向量表中的位用于表示相应分词在所述模式串中是否出现;

根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配;

所述分配所述编码后的模式串对应的位向量表,包括:

分配所述目标子串对应的位向量表;

所述根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配,包括:

根据所述目标子串对应的位向量表,基于位并行算法对所述目标子串与所述文本串进行匹配,得到匹配串在所述文本串中的位置;

根据所述匹配串在所述文本串中的位置,查询所述文本串是否命中所述目标子串对应的模式串。

2.根据权利要求1所述的方法,其特征在于,所述根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配之前,所述方法还包括:

对所述文本串进行分词,得到所述文本串对应的第二分词集合;

对所述第二分词集合中的各分词进行编码,得到编码后的文本串;

所述根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配,包括:

根据所述位向量表,基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

3.根据权利要求1所述的方法,其特征在于,所述在所述各个子串中,确定目标子串,包括:

在所述各个子串中,确定词频最小的子串为目标子串。

4.根据权利要求1所述的方法,其特征在于,所述根据所述匹配串在所述文本串中的位置,查询所述文本串是否命中所述目标子串对应的模式串,包括:

若所述目标子串存在碰撞模式串,则分别计算每个碰撞模式串的词频;

根据所述匹配串在所述文本串中的位置,按照碰撞模式串词频从大到小的顺序,依次查询所述文本串是否命中所述碰撞模式串。

5.根据权利要求1所述的方法,其特征在于,若所述待匹配的模式串的个数n大于1,所述分配所述编码后的模式串对应的位向量表,包括:

确定位向量表的分组个数m;

将n个编码后的模式串分配到包含m个分组的位向量表中。

6.根据权利要求5所述的方法,其特征在于,所述将n个编码后的模式串分配到包含m个分组的位向量表中,包括:

确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合;其中,i的取值为1～n,j的取值为1～m;

根据所述误判字符串集合中各误判字符串的词频,确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益;

将所述第i个编码后的模式串分配到第一损失增益最小的分组,直到第n个编码后的模式串分配完成。

7.一种模式匹配装置,其特征在于,所述装置包括:

第一分词模块,用于对待匹配的模式串进行分词,得到所述模式串对应的第一分词集合;

第一编码模块,用于对所述第一分词集合中的各分词进行编码,得到编码后的模式串;

分配模块,用于分配所述编码后的模式串对应的位向量表,所述位向量表中的位用于表示相应分词在所述模式串中是否出现;

模式匹配模块,用于根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配;

子串划分模块,用于在所述编码后的模式串的长度大于预设的分组长度时,按照所述分组长度,对所述编码后的模式串进行划分,得到划分后的各个子串;

目标确定模块,用于在所述各个子串中,确定目标子串;

所述分配模块,具体用于分配所述目标子串对应的位向量表;

所述模式匹配模块,包括:

匹配子模块,用于根据所述目标子串对应的位向量表,基于位并行算法对所述目标子串与所述文本串进行匹配,得到匹配串在所述文本串中的位置;

查询子模块,用于根据所述匹配串在所述文本串中的位置,查询所述文本串是否命中所述目标子串对应的模式串。

8.根据权利要求7所述的装置,其特征在于,所述装置还包括:

第二分词模块,用于对所述文本串进行分词,得到所述文本串对应的第二分词集合;

第二编码模块,用于对所述第二分词集合中的各分词进行编码,得到编码后的文本串;

所述模式匹配模块,具体用于根据所述位向量表,基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

9.根据权利要求7所述的装置,其特征在于,所述目标确定模块,具体用于在所述各个子串中,确定词频最小的子串为目标子串。

10.根据权利要求7所述的装置,其特征在于,所述查询子模块,包括:

词频计算单元,用于若所述目标子串存在碰撞模式串,则分别计算每个碰撞模式串的词频;

查询比较单元,用于根据所述匹配串在所述文本串中的位置,按照碰撞模式串词频从大到小的顺序,依次查询所述文本串是否命中所述碰撞模式串。

11.根据权利要求7所述的装置,其特征在于,若所述待匹配的模式串的个数n大于1,所述分配模块,包括:

分组确定子模块,用于确定位向量表的分组个数m;

模式串分配子模块,用于将n个编码后的模式串分配到包含m个分组的位向量表中。

12.根据权利要求11所述的装置,其特征在于,所述模式串分配子模块,包括:

第一计算单元,用于确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合;其中,i的取值为1～n,j的取值为1～m;

第二计算单元,用于根据所述误判字符串集合中各误判字符串的词频,确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益;

模式串分配单元,用于将所述第i个编码后的模式串分配到第一损失增益最小的分组,直到第n个编码后的模式串分配完成。

13.一种用于数据处理的装置,其特征在于,包括有存储器,以及一个或者一个以上的程序,其中一个或者一个以上程序存储于存储器中,且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令:

对待匹配的模式串进行分词,得到所述模式串对应的第一分词集合;

对所述第一分词集合中的各分词进行编码,得到编码后的模式串;

分配所述编码后的模式串对应的位向量表,所述位向量表中的位用于表示相应分词在所述模式串中是否出现;

根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配;

所述装置还经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令:

在所述编码后的模式串的长度大于预设的分组长度时,按照所述分组长度,对所述编码后的模式串进行划分,得到划分后的各个子串;

在所述各个子串中,确定目标子串;

所述分配所述编码后的模式串对应的位向量表,包括:

分配所述目标子串对应的位向量表;

所述根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配,包括:

根据所述目标子串对应的位向量表,基于位并行算法对所述目标子串与所述文本串进行匹配,得到匹配串在所述文本串中的位置;

根据所述匹配串在所述文本串中的位置,查询所述文本串是否命中所述目标子串对应的模式串。

14.根据权利要求13所述的装置,其特征在于,所述装置还经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令:

对所述文本串进行分词,得到所述文本串对应的第二分词集合;

对所述第二分词集合中的各分词进行编码,得到编码后的文本串;

所述根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配,包括:

根据所述位向量表,基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

15.根据权利要求13所述的装置,其特征在于,所述在所述各个子串中,确定目标子串,包括:

在所述各个子串中,确定词频最小的子串为目标子串。

16.根据权利要求13所述的装置，其特征在于，所述根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串，包括：

若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，依次查询所述文本串是否命中所述碰撞模式串。

17.根据权利要求13所述的装置，其特征在于，若所述待匹配的模式串的个数n大于1，所述分配所述编码后的模式串对应的位向量表，包括：

确定位向量表的分组个数m；

将n个编码后的模式串分配到包含m个分组的位向量表中。

18.根据权利要求17所述的装置，其特征在于，所述将n个编码后的模式串分配到包含m个分组的位向量表中，包括：

确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

19.一种机器可读介质，其上存储有指令，当由一个或多个处理器执行时，使得装置执行如权利要求1至6中一个或多个所述的模式匹配方法。

一种模式匹配方法、装置和用于模式匹配的装置

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种模式匹配方法、装置和用于模式匹配的装置。

背景技术

[0002] 字符串匹配又称模式匹配，是广泛应用于信息检索、入侵检测、计算生物学、搜索引擎、数据压缩等领域的一项关键技术。所谓模式匹配，指的是查找某个模式串P＝p1p2...pm在文本串T＝t1t2...tn中的所有出现位置。

[0003] 位并行算法是目前较为常用的一种模式匹配算法，位并行算法包括shift-and(移位-与)、shift-or(移位-或)、BNDM(Backward Nondeterministic Dawg Matching，后缀自动机匹配)。通常，位并行算法会在计算机缓存中维护一个位向量表，简称B表。B表可以理解成一个n×m的0/1矩阵，表中的0/1用于表示相应的字符是否在模式串中出现。

[0004] 然而，B表中每一位表示一个字符，导致B表占用大量内存空间。此外，在实际应用中，由于语义的特殊性，并非每个匹配到的结果都有意义。例如，文本串为“刘德华为什么那么帅”，模式串为“华为”，则会匹配成功，但这种匹配并没有意义，因为该文本串与“华为”没有语义关联关系，导致存在无意义的匹配结果。

发明内容

[0005] 本发明实施例提供一种模式匹配方法、装置和用于模式匹配的装置，可以提高B表的利用率，以及提高模式匹配的效率。

[0006] 为了解决上述问题，本发明实施例公开了一种模式匹配方法，所述方法包括：

[0007] 对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0008] 对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0009] 分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0010] 根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0011] 另一方面，本发明实施例公开了一种模式匹配装置，所述装置包括：

[0012] 第一分词模块，用于对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0013] 第一编码模块，用于对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0014] 分配模块，用于分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0015] 模式匹配模块，用于根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0016] 再一方面，本发明实施例公开了一种用于模式匹配的装置，包括有存储器，以及一

个或者一个以上的程序，其中一个或者一个以上程序存储于存储器中，且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：

[0017] 对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0018] 对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0019] 分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0020] 根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0021] 又一方面，本发明实施例公开了一种机器可读介质，其上存储有指令，当由一个或多个处理器执行时，使得装置执行如前述一个或多个所述的模式匹配方法。

[0022] 本发明实施例包括以下优点：

[0023] 本发明实施例在模式匹配之前，对模式串进行分词，得到所述模式串对应的第一分词序列，并且对所述第一分词序列中的各分词进行编码，得到编码后的模式串；以及分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；进而根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0024] 首先，本发明实施例对模式串进行分词后再编码，不仅可以修正语义错误，过滤无意义的匹配结果，而且分词后的编码可以压缩B表的值域长度，缩减占用缓存的空间，提高B表的利用率。

[0025] 此外，本发明实施例通过子串的词频选择目标子串，可以减少命中目标子串时产生碰撞模式串的概率，进而可以提高后续的查询效率。

[0026] 再者，在目标子串存在碰撞模式串的情况下，本发明实施例基于模式串的词频确定碰撞模式串的查询顺序，以尽可能地减少查询次数，提高查询效率。

附图说明

[0027] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0028] 图1是本发明的一种模式匹配方法实施例的步骤流程图；

[0029] 图2是本发明的一种模式匹配装置实施例的结构框图；

[0030] 图3是本发明的一种用于模式匹配的装置800的框图；及

[0031] 图4是本发明的一些实施例中服务器的结构示意图。

具体实施方式

[0032] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0033] 方法实施例

[0034] 参照图1,示出了本发明的一种模式匹配方法实施例的步骤流程图,具体可以包括如下步骤:

[0035] 步骤101、对待匹配的模式串进行分词,得到所述模式串对应的第一分词集合;

[0036] 步骤102、对所述第一分词集合中的各分词进行编码,得到编码后的模式串;

[0037] 步骤103、分配所述编码后的模式串对应的位向量表,所述位向量表中的位用于表示相应分词在所述模式串中是否出现;

[0038] 步骤104、根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0039] 本发明实施例的模式匹配方法可适用于电子设备,所述电子设备包括但不限于:服务器、智能手机、录音笔、平板电脑、电子书阅读器、MP3(动态影像专家压缩标准音频层面3,Moving Picture Experts Group Audio Layer III)播放器、MP4(动态影像专家压缩标准音频层面4,Moving Picture Experts Group Audio Layer IV)播放器、膝上型便携计算机、车载电脑、台式计算机、机顶盒、智能电视机、可穿戴设备等等。

[0040] 本发明实施例的模式匹配方法可用于单模匹配或多模匹配,其中,单模匹配指在一个文本串中匹配一个模式串;多模匹配指在一个文本串中匹配多个模式串。

[0041] 在对文本串和模式串进行匹配之前,本发明实施例对模式串进行分词处理,得到所述模式串对应的第一分词集合。

[0042] 在本发明实施例的一个中文示例中,假设待匹配的文本串为“刘德华为什么那么帅”,待匹配的模式串包括:“为什么那么帅”、“华为”共两个模式串。首先,对待匹配的两个模式串分别进行分词,根据分词结果可以得到第一分词集合{“为什么”、“那么”、“帅”、“华为”}。然后,对第一分词集合中的各分词进行编码,得到编码后的模式串。例如,将“为什么”编码为C,“那么”编码为D,“帅”编码为E,“华为”编码为F,则编码后的两个模式串分别为:CDE和F。

[0043] 在本发明的一种可选实施例中,步骤104根据所述位向量表,基于位并行算法对所述模式串与待匹配的文本串进行匹配之前,所述方法还可以包括:

[0044] 步骤S11、对所述文本串进行分词,得到所述文本串对应的第二分词集合;

[0045] 步骤S12、对所述第二分词集合中的各分词进行编码,得到编码后的文本串。

[0046] 在模式匹配系统读入文本串之后,可以对文本串进行分词,得到所述文本串对应的第二分词集合。例如,在上述示例中,可以对文本串“刘德华为什么那么帅”进行分词,得到第二分词集合{“刘德华”、“为什么”、“那么”、“帅”},然后对第二分词集合中的各分词进行编码,可以得到编码后的文本串。例如,将“为什么”编码为C,“那么”编码为D,“帅”编码为E,“刘德华”编码为A,则编码后的文本串为:ACDE。其中,第一分词集合和第二分词集合在编码过程中,可以采用相同的编码规则,使得相同的分词可以对应相同的编码。

[0047] 通常,位并行算法会在计算机缓存中维护一个位向量表,简称B表,B表的长度不超过w,w为计算机一次处理的位长(一个机器字长,如64位)。B表可以是一个n×m的0/1矩阵,表中的0/1用于表示相应位置的字符是否在模式串中出现,n为模式串的值域长度。

[0048] 以64位的位向量表为例,在单模匹配(待匹配的模式串的个数n=1)的情况下,可以直接将待匹配的模式串分配至64位的向量表中。而在多模匹配(待匹配的模式串的个数n大于1)的情况下,需要对位向量表进行分组使用,且所有分组的位长之和不超过w(64位)。

对于分组个数为m的位向量表，首先将n个编码后的模式串分配到包含m个分组的位向量表中，然后，在一个机器字的多个分组中，同时对多个模式串执行shift-and算法的位操作。将n个编码后的模式串分配到包含m个分组的位向量表中，可以存在如下两种情况，在n小于或等于m的情况下，每个模式串可以单独分配一个分组；在n大于m的情况下，可能会将多个模式串分配到同一个分组中。为便于描述，本发明实施例中均以n小于或等于m的情况为例进行说明，在n大于m的情况下，只需对一个分组中的多个模式串再进行比较以确认到底命中的是哪个模式串即可。

[0049] 本发明实施例步骤102在对所述第一分词集合中的各分词进行编码，得到编码后的模式串之后，可以对编码后的模式串分配位向量表(B表)，由此，该编码后的模式串对应的B表中的位可用于表示相应分词在待匹配的模式串中是否出现。

[0050] 步骤104根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，具体可以包括：根据所述位向量表，基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

[0051] 例如，在上述示例中，在文本串“刘德华为什么那么帅”中匹配模式串“为什么那么帅”和“华为”的问题被转换为：在编码后的文本串“ACDE”中匹配编码后的模式串“CDE”和“F”的问题。可以看出，“ACDE”中并不存在“F”，因此，最终得到的匹配串包括“为什么那么帅”，但是不包括“华为”，本发明实施例基于分词后的模式串进行编码，可以过滤无意义的匹配结果。

[0052] 此外，本发明实施例基于分词后的模式串分配B表，可以减小B表的值域长度，提高B表的利用率。例如，在上述示例中，对于模式串“为什么那么帅”和“华为”，按照传统方式分配B表，B表中每一位表示一个字符，B表的值域长度通常为2的指数级。然而，本发明实施例对模式串“为什么那么帅”和“华为”进行分词，将分词“为什么”编码为C，“那么”编码为D，“帅”编码为E，“华为”编码为F，模式串中未出现的分词统一编码为G。由此，B表的值域长度仅为5，大大减小了B表的值域长度，进而可以提高B表的利用率。参见表1，示出了本发明实施例对模式串“为什么那么帅”分配B表的一个示例。

[0053] 表1

[0054]

C	001

[0055]

D	010
E	100
G	000

[0056] 如表1所示，对模式串“为什么那么帅”进行分词及编码后得到编码后的模式串为“CDE”，其中，“C”为分词“为什么”对应的编码，如表1所示，第一行为编码“C”对应的位掩码“001”。位掩码中的0/1用于表示相应位置的字符是否在模式串中出现，需要说明的是，在本发明实施例中，模式串中的字符表示顺序与位掩码中的字符表示顺序相反。例如，在“001”中，从右向左第1位为1，其余位为0，表示在编码后的模式串“CDE”中，从左向右第1位为“C”。同理，如表1所示，第二行为编码“D”对应的位掩码“010”。在“010”中，从右向左第2位为1，其

余位为0，表示在编码后的模式串“CDE”中，从左向右第2位为“D”。第三行为编码“E”对应的位掩码“100”。在“100”中，从右向左第3位为1，其余位为0，表示在编码后的模式串“CDE”中，从左向右第3位为“E”。第四行的位掩码“000”表示编码后的模式串“CDE”中未出现的字符，也即表示模式串“为什么那么帅”中未出现的其他分词。

[0057] 需要说明的是，上述将“为什么”编码为“C”，“那么”编码为“D”，“帅”编码为“E”，“华为”编码为“F”，其它词统一编码为“G”的编码方式仅作为本发明实施例的一种应用示例，本发明实施例对具体的编码方式不加以限制。

[0058] 例如，在多模匹配的情况下，对于待匹配的n个模式串，首先将这n个模式串进行分词，得到分词后的分词集合。然后对分词集合中的各分词进行编码，每一个分词对应一个编码。假设对所述n个模式串进行分词得到的分词集合中有100个分词，从编码0开始对这100个分词进行编码，那么这100个分词对应的编码为从0至99，每一个编码对应一个分词，如第一个分词对应的编码为0，第二个分词对应的编码为1，以此类推，第100个分词对应的编码为99，以及将这100个分词中未出现的分词编码为100。接下来根据上述分词对应的编码对待匹配的n个模式串进行编码，可以得到编码后的n个模式串。最后，对编码后的n个模式串分配B表。

[0059] 本发明实施例对模式串进行分词后再编码，不仅可以修正语义错误，过滤无意义的匹配结果，而且分词后的编码可以压缩B表的值域长度，缩减占用缓存的空间，提高B表的利用率。

[0060] 在本发明的一种可选实施例中，步骤103分配所述编码后的模式串对应的位向量表之前，所述方法还可以包括：

[0061] 步骤S21、在所述编码后的模式串的长度大于预设的分组长度时，按照所述分组长度，对所述编码后的模式串进行划分，得到划分后的各子串；

[0062] 步骤S22、在所述各子串中，确定目标子串；

[0063] 基于此，步骤103分配所述编码后的模式串对应的位向量表，具体可以包括：分配所述目标子串对应的位向量表。

[0064] 在具体应用中，多模匹配包括多个待匹配的模式串，因此通常需要对B表进行分组使用，且所有分组的位长之和不超过w(计算机一次处理的位长，如一个机器字长)。在位并行算法中，当一个模式串的长度h大于B表的分组长度M时，需要对模式串进行截断。其中，分组长度M可以根据实际需要预先设置。一般的做法是截取模式串的前M个字符的子串或后M个字符的子串来代替模式串。

[0065] 在本发明实施例的一个英文示例中，对于文本串“ababfdeabc”，假设待匹配的模式串包括“abcde”和“abfde”共两个模式串，且将模式串“abcde”和“abfde”分配在B表的同一个分组中，假设该分组的长度M为2bit，模式串“abcde”和“abfde”的长度均大于分组长度，因此需要对模式串“abcde”和“abfde”分别进行截断，假设采用截取前2个字符的方式进行截断，则这两个模式串都是截取子串“ab”作为目标子串，也即使用目标子串“ab”代替模式串“abcde”，以及使用目标子串“ab”代替模式串“abfde”。此时，只需分配目标子串“ab”对应的B表。

[0066] 可选地，步骤104根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，具体可以包括：

[0067] 步骤S31、根据所述目标子串对应的位向量表，基于位并行算法对所述目标子串与所述文本串进行匹配，得到匹配串在所述文本串中的位置；

[0068] 步骤S32、根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串。

[0069] 在上述示例中，对于模式串s1＝“abcde”和s2＝“abfde”，都使用相同的目标子串“ab”，分配得到的B表如表2所示。

[0070] 表2

[0071]

a	01
b	10
c	00

[0072] 根据表2所示的B表，基于位并行算法对该目标子串“ab”与文本串“ababfdeabc”进行匹配，可以得到3个匹配串“ab”，这三个匹配串在文本串“ababfdeabc”中的位置分别为pos＝0，pos＝2，pos＝7。

[0073] 在本发明实施例中，假设一个目标子串对应k个模式串，k为大于1的整数，在使用该目标子串对待匹配的文本串进行匹配的过程中，如果该文本串命中该目标子串，也即该文本串中存在该目标子串时，会产生所述k个模式串的碰撞，本发明实施例称所述k个模式串为碰撞模式串。

[0074] 例如，在上述示例中，目标子串“ab”对应两个模式串(k＝2)，分别为模式串“abcde”和模式串“abfde”。在使用该目标子串“ab”与文本串“ababfdeabc”进行匹配的过程中，在读入该文本串第二个字符“b”时，可以发现该文本串命中该目标子串“ab”。由于目标子串“ab”同时对应模式串“abcde”和“abfde”，因此模式串“abcde”和“abfde”产生碰撞，称“abcde”和“abfde”为碰撞模式串。此时，无法确定模式串“abcde”和“abfde”中到底是哪个模式串和文本串“ababfdeabc”相匹配。

[0075] 可以看出，在使用目标子串代替模式串的情况下，由于目标子串仅表示模式串的部分内容，因此，文本串命中目标子串，仅表示该文本串存在命中该目标子串对应的模式串的可能性，需对该目标子串产生的所有碰撞模式串进行遍历，根据匹配串在该文本串中的位置，进一步查询该文本串是否命中每个碰撞模式串。

[0076] 例如，目标子串“ab”产生碰撞模式串“abcde”和“abfde”，假设查询顺序为先查询模式串“abcde”，再查询模式串“abfde”。首先，比较文本串“ababfdeabc”中从位置pos＝0开始，长度为5的字符串“ababf”与模式串“abcde”是否相匹配，结果为不匹配。然后，比较文本串“ababfdeabc”中从位置pos＝2开始，长度为5的字符串“abfde”与模式串“abcde”是否相匹配，结果为不匹配。最后，比较文本串“ababfdeabc”中从位置pos＝7开始，长度为5的字符串与模式串“abcde”是否相匹配，由于从位置pos＝7开始到字符串最后一个字符的长度仅为3，因此不匹配。

[0077] 在对模式串“abcde”查询完成之后，发现文本串“ababfdeabc”中不存在模式串“abcde”，也即，文本串“ababfdeabc”未命中模式串“abcde”，此时，再按照上述查询过程，查询文本串“ababfdeabc”是否命中模式串“abfde”。查询结果为文本串“ababfdeabc”命中模式串“abfde”，且模式串“abfde”在文本串“ababfdeabc”中的位置为pos＝2。

[0078] 可以看出，在上述示例中，截取模式串“abcde”和“abfde”的前两个字符“ab”作为

目标子串代替模式串，导致匹配过程中产生碰撞模式串“abcde”和“abfde”。在实际应用中，模式串的数量越多，产生碰撞模式串的数量可能也越多，后续需要进一步对每个碰撞模式串进行查询比较，以确定文本串是否真的命中该碰撞模式串，导致匹配效率较低。

[0079] 在本发明的一种可选实施例中，步骤S32根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串，具体可以包括：

[0080] 步骤S41、若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

[0081] 步骤S42、根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，依次查询所述文本串是否命中所述碰撞模式串。

[0082] 在上述示例中，模式串包括“abcde”和“abfde”，若截取模式串“abcde”和“abfde”的前两个字符“ab”作为目标子串代替模式串，在与文本串“ababfdeabc”进行匹配的过程中，目标子串“ab”会存在碰撞模式串“abcde”和“abfde”。

[0083] 通过上述示例可以看出，在目标子串存在碰撞模式串的情况下，在文本串中查询碰撞模式串的顺序将影响查询的效率。为了提高查询效率，本发明实施例计算碰撞模式串中每个模式串的词频。

[0084] 具体地，可以基于给定的语料库，分别计算碰撞模式串中每个模式串的词频，如分别计算模式串“abcde”和“abfde”的词频，假设F(abcde)＝10，F(abfde)＝20，则将模式串“abcde”和“abfde”按照词频从大到小挂入链表或顺序存储，按照词频顺序进行查询，也即先查询模式串“abfde”，再查询模式串“abcde”。

[0085] 由于碰撞模式串的词频可以根据语料库统计得到，说明在通常情况下，模式串“abfde”出现的概率比模式串“abcde”出现的概率高，因此，先搜索词频较高的模式串，命中的概率较高，可以减少后续查询的次数，提高模式匹配的效率。

[0086] 在本发明的一种可选实施例中，步骤S22在所述各个子串中，确定目标子串，具体可以包括：在所述各个子串中，确定词频最小的子串为目标子串。

[0087] 为了尽可能地减少目标子串产生碰撞模式串的概率，本发明实施例根据词频确定目标子串。具体地，当模式串的长度h超过B表的组长M时，按B表的组长M对模式串依次进行划分，得到模式串的各子串，并分别计算划分后得到的各子串的词频，选择词频最小的子串作为目标子串。

[0088] 例如，在上述示例中，对于模式串“abcde”，长度h＝5，假设该模式串分配的B表分组的组长M＝2。按B表的组长M对该模式串进行划分，得到该模式串的各子串为：“ab”、“bc”、“cd”、“de”、“bf”、“fd”、“de”。分别计算各子串的词频F，假设F(ab)＝2，F(bc)＝1，F(cd)＝1，F(de)＝2，F(bf)＝1，F(fd)＝1，对于模式串“abcde”，子串“bc”和子串“cd”的词频最小，因此可以将子串“bc”或者“cd”作为模式串“abcde”的目标子串。同理，对于模式串“abfde”，子串“cd”和子串“bf”的词频最小，因此可以将子串“cd”或子串“bf”作为模式串“abfde”的目标子串。

[0089] 在一个示例中，假设将子串“bc”作为模式串“abcde”的目标子串，以及将子串“bf”作为模式串“abfde”的目标子串，分配得到的B表如表3所示。

[0090] 表3

[0091]

a	00
b	01

c	10
d	00
e	00
f	10
g	00

[0092] 根据表3对文本串进行匹配的过程中，在文本串命中目标子串“bc”或者命中目标子串“bf”时，模式串“abcde”和模式串“abfde”不会产生碰撞。可以看出，本发明实施例通过子串的词频选择目标子串，可以减少命中目标子串时产生碰撞模式串的概率，进而可以提高后续的查询效率。

[0093] 可以理解，本发明实施例对计算子串词频的具体方式不加以限制。例如，可以根据当前文本串所在应用环境下的文档进行计算，或者也可以根据大数据进行计算等。

[0094] 在本发明的一种可选实施例中，步骤103中，若所述待匹配的模式串的个数n大于1，所述分配所述编码后的模式串对应的位向量表，具体可以包括：

[0095] 步骤S51、确定位向量表的分组个数m；

[0096] 步骤S52、将n个编码后的模式串分配到包含m个分组的位向量表中。

[0097] 在多模匹配(待匹配的模式串的个数n大于1)的情况下，需要对位向量表进行分组使用，且所有分组的位长之和不超过w(64位)。对于分组个数为m的位向量表，首先将n个编码后的模式串分配到包含m个分组的位向量表中，然后，在一个机器字的多个分组中，同时对多个模式串执行shift-and算法的位操作。将n个编码后的模式串分配到包含m个分组的位向量表中，可以存在如下两种情况，在n小于或等于m的情况下，每个模式串可以单独分配一个分组；在n大于m的情况下，可能会将多个模式串分配到同一个分组中。

[0098] 例如，在一个分组中分配有5个模式串，在读取文本串的某个字符时，该字符可能会同时命中这5个模式串中的多个，产生模式串的碰撞，需要进一步比较到底命中的是哪个模式串，增加比较操作成本，影响匹配的效率。

[0099] 在本发明的一种可选实施例中，所述将n个编码后的模式串分配到包含m个分组的位向量表中，具体可以包括：

[0100] 步骤S61、确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

[0101] 步骤S62、根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

[0102] 步骤S63、将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

[0103] 为了尽可能地减少模式串的碰撞，提高模式匹配的效率，在多模匹配过程中，在给定分组数与分组长度的情况下，本发明实施例基于贪心原则按最小损失增益将n个模式串分配到m个分组中。

[0104] 综上，本发明实施例在模式匹配之前，对模式串进行分词，得到所述模式串对应的第一分词集合，并且对所述第一分词集合中的各分词进行编码，得到编码后的模式串；以及分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；进而根据所述位向量表，基于位并行算法对所述模式串与待匹配的文

本串进行匹配。

[0105] 首先，本发明实施例对模式串进行分词后再编码，不仅可以修正语义错误，过滤无意义的匹配结果，而且分词后的编码可以压缩B表的值域长度，缩减占用缓存的空间，提高B表的利用率。

[0106] 此外，本发明实施例通过子串的词频选择目标子串，可以减少命中目标子串时产生碰撞模式串的概率，进而可以提高后续的查询效率。

[0107] 再者，在目标子串存在碰撞模式串的情况下，本发明实施例基于模式串的词频确定碰撞模式串的查询顺序，以尽可能地减少查询次数，提高查询效率。

[0108] 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明实施例并不受所描述的动作顺序的限制，因为依据本发明实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明实施例所必须的。

[0109] 装置实施例

[0110] 参照图2，示出了本发明的一种模式匹配装置实施例的结构框图，所述装置具体可以包括：

[0111] 第一分词模块201，用于对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0112] 第一编码模块202，用于对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0113] 分配模块203，用于分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0114] 模式匹配模块204，用于根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0115] 可选地，所述装置还包括：

[0116] 第二分词模块，用于对所述文本串进行分词，得到所述文本串对应的第二分词集合；

[0117] 第二编码模块，用于对所述第二分词集合中的各分词进行编码，得到编码后的文本串；

[0118] 所述模式匹配模块，具体用于根据所述位向量表，基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

[0119] 可选地，所述装置还包括：

[0120] 子串划分模块，用于在所述编码后的模式串的长度大于预设的分组长度时，按照所述分组长度，对所述编码后的模式串进行划分，得到划分后的各子串；

[0121] 目标确定模块，用于在所述各子串中，确定目标子串；

[0122] 所述分配模块，具体用于分配所述目标子串对应的位向量表；

[0123] 所述模式匹配模块，包括：

[0124] 匹配子模块，用于根据所述目标子串对应的位向量表，基于位并行算法对所述目标子串与所述文本串进行匹配，得到匹配串在所述文本串中的位置；

[0125] 查询子模块，用于根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串。

[0126] 可选地，所述目标确定模块，具体用于在所述各个子串中，确定词频最小的子串为目标子串。

[0127] 可选地，所述查询子模块，包括：

[0128] 词频计算单元，用于若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

[0129] 查询比较单元，用于根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，依次查询所述文本串是否命中所述碰撞模式串。

[0130] 可选地，若所述待匹配的模式串的个数n大于1，所述分配模块，包括：

[0131] 分组确定子模块，用于确定位向量表的分组个数m；

[0132] 模式串分配子模块，用于将n个编码后的模式串分配到包含m个分组的位向量表中。

[0133] 可选地，所述分组分配子模块，包括：

[0134] 第一计算单元，用于确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

[0135] 第二计算单元，用于根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

[0136] 模式串分配单元，用于将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

[0137] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0138] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0139] 关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0140] 本发明实施例提供了一种用于模式匹配的装置，包括有存储器，以及一个或者一个以上的程序，其中一个或者一个以上程序存储于存储器中，且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：对模式串进行分词，得到所述模式串对应的第一分词集合；对所述第一分词集合中的各分词进行编码，得到编码后的模式串；分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0141] 图3是根据一示例性实施例示出的一种用于模式匹配的装置800的框图。例如，装置800可以是移动电话，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理等。

[0142] 参照图3，装置800可以包括以下一个或多个组件：处理组件802，存储器804，电源组件806，多媒体组件808，音频组件810，输入/输出(I/O)的接口812，传感器组件814，以及通信组件816。

[0143] 处理组件802通常控制装置800的整体操作，诸如与显示，电话呼叫，数据通信，相机操作和记录操作相关联的操作。处理元件802可以包括一个或多个处理器820来执行指令，以完成上述的方法的全部或部分步骤。此外，处理组件802可以包括一个或多个模块，便于处理组件802和其他组件之间的交互。例如，处理组件802可以包括多媒体模块，以方便多媒体组件808和处理组件802之间的交互。

[0144] 存储器804被配置为存储各种类型的数据以支持在设备800的操作。这些数据的示例包括用于在装置800上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。存储器804可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器(SRAM)，电可擦除可编程只读存储器(EEPROM)，可擦除可编程只读存储器(EPROM)，可编程只读存储器(PROM)，只读存储器(ROM)，磁存储器，快闪存储器，磁盘或光盘。

[0145] 电源组件806为装置800的各种组件提供电力。电源组件806可以包括电源管理系统，一个或多个电源，及其他与为装置800生成、管理和分配电力相关联的组件。

[0146] 多媒体组件808包括在所述装置800和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中，多媒体组件808包括一个前置摄像头和/或后置摄像头。当设备800处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0147] 音频组件810被配置为输出和/或输入音频信号。例如，音频组件810包括一个麦克风(MIC)，当装置800处于操作模式，如呼叫模式、记录模式和语音信息处理模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器804或经由通信组件816发送。在一些实施例中，音频组件810还包括一个扬声器，用于输出音频信号。

[0148] I/O接口812为处理组件802和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

[0149] 传感器组件814包括一个或多个传感器，用于为装置800提供各个方面的状态评估。例如，传感器组件814可以检测到设备800的打开/关闭状态，组件的相对定位，例如所述组件为装置800的显示器和小键盘，传感器组件814还可以检测装置800或装置800一个组件的位置改变，用户与装置800接触的存在或不存在，装置800方位或加速/减速和装置800的温度变化。传感器组件814可以包括接近传感器，被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件814还可以包括光传感器，如CMOS或CCD图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件814还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

[0150] 通信组件816被配置为便于装置800和其他设备之间有线或无线方式的通信。装置800可以接入基于通信标准的无线网络，如WiFi，2G或3G，或它们的组合。在一个示例性实施例中，通信组件816经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。

在一个示例性实施例中，所述通信组件816还包括近场通信(NFC)模块，以促进短程通信。例如，在NFC模块可基于射频信息处理(RFID)技术，红外数据协会(IrDA)技术，超宽带(UWB)技术，蓝牙(BT)技术和其他技术来实现。

[0151] 在示例性实施例中，装置800可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

[0152] 在示例性实施例中，还提供了一种包括指令的非临时性计算机可读存储介质，例如包括指令的存储器804，上述指令可由装置800的处理器820执行以完成上述方法。例如，所述非临时性计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0153] 图4是本发明的一些实施例中服务器的结构示意图。该服务器1900可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器(central processing units,CPU)1922(例如，一个或一个以上处理器)和存储器1932，一个或一个以上存储应用程序1942或数据1944的存储介质1930(例如一个或一个以上海量存储设备)。其中，存储器1932和存储介质1930可以是短暂存储或持久存储。存储在存储介质1930的程序可以包括一个或一个以上模块(图示没标出)，每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器1922可以设置为与存储介质1930通信，在服务器1900上执行存储介质1930中的一系列指令操作。

[0154] 服务器1900还可以包括一个或一个以上电源1926，一个或一个以上有线或无线网络接口1950，一个或一个以上输入输出接口1958，一个或一个以上键盘1956，和/或，一个或一个以上操作系统1941，例如Windows ServerTM,Mac OS XTM,UnixTM,LinuxTM,FreeBSDTM等等。

[0155] 一种非临时性计算机可读存储介质，当所述存储介质中的指令由装置(服务器或者终端)的处理器执行时，使得装置能够执行图1所示的模式匹配方法。

[0156] 一种非临时性计算机可读存储介质，当所述存储介质中的指令由装置(服务器或者终端)的处理器执行时，使得装置能够执行一种模式匹配方法，所述方法包括：对模式串进行分词，得到所述模式串对应的第一分词集合；对所述第一分词集合中的各分词进行编码，得到编码后的模式串；分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0157] 本发明实施例公开了A1、一种模式匹配方法，包括：

[0158] 对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0159] 对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0160] 分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0161] 根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0162] A2、根据A1所述的方法，所述根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配之前，所述方法还包括：

[0163] 对所述文本串进行分词，得到所述文本串对应的第二分词集合；

[0164] 对所述第二分词集合中的各分词进行编码，得到编码后的文本串；

[0165] 所述根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，包括：

[0166] 根据所述位向量表，基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

[0167] A3、根据A1所述的方法，所述分配所述编码后的模式串对应的位向量表之前，所述方法还包括：

[0168] 在所述编码后的模式串的长度大于预设的分组长度时，按照所述分组长度，对所述编码后的模式串进行划分，得到划分后的各子串；

[0169] 在所述各子串中，确定目标子串；

[0170] 所述分配所述编码后的模式串对应的位向量表，包括：

[0171] 分配所述目标子串对应的位向量表；

[0172] 所述根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，包括：

[0173] 根据所述目标子串对应的位向量表，基于位并行算法对所述目标子串与所述文本串进行匹配，得到匹配串在所述文本串中的位置；

[0174] 根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串。

[0175] A4、根据A3所述的方法，所述在所述各个子串中，确定目标子串，包括：

[0176] 在所述各个子串中，确定词频最小的子串为目标子串。

[0177] A5、根据A3所述的方法，所述根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串，包括：

[0178] 若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

[0179] 根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，依次查询所述文本串是否命中所述碰撞模式串。

[0180] A6、根据A1所述的方法，若所述待匹配的模式串的个数n大于1，所述分配所述编码后的模式串对应的位向量表，包括：

[0181] 确定位向量表的分组个数m；

[0182] 将n个编码后的模式串分配到包含m个分组的位向量表中。

[0183] A7、根据A6所述的方法，所述将n个编码后的模式串分配到包含m个分组的位向量表中，包括：

[0184] 确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

[0185] 根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

[0186] 将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

[0187] 本发明实施例公开了B8、一种模式匹配装置，所述装置包括：

[0188] 第一分词模块，用于对待匹配的模式串进行分词，得到所述模式串对应的第一分

词集合；

[0189] 第一编码模块，用于对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0190] 分配模块，用于分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0191] 模式匹配模块，用于根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0192] B9、根据B8所述的装置，所述装置还包括：

[0193] 第二分词模块，用于对所述文本串进行分词，得到所述文本串对应的第二分词集合；

[0194] 第二编码模块，用于对所述第二分词集合中的各分词进行编码，得到编码后的文本串；

[0195] 所述模式匹配模块，具体用于根据所述位向量表，基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

[0196] B10、根据B8所述的装置，所述装置还包括：

[0197] 子串划分模块，用于在所述编码后的模式串的长度大于预设的分组长度时，按照所述分组长度，对所述编码后的模式串进行划分，得到划分后的各子串；

[0198] 目标确定模块，用于在所述各子串中，确定目标子串；

[0199] 所述分配模块，具体用于分配所述目标子串对应的位向量表；

[0200] 所述模式匹配模块，包括：

[0201] 匹配子模块，用于根据所述目标子串对应的位向量表，基于位并行算法对所述目标子串与所述文本串进行匹配，得到匹配串在所述文本串中的位置；

[0202] 查询子模块，用于根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串。

[0203] B11、根据B10所述的装置，所述目标确定模块，具体用于在所述各个子串中，确定词频最小的子串为目标子串。

[0204] B12、根据B10所述的装置，所述查询子模块，包括：

[0205] 词频计算单元，用于若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

[0206] 查询比较单元，用于根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，依次查询所述文本串是否命中所述碰撞模式串。

[0207] B13、根据B8所述的装置，若所述待匹配的模式串的个数n大于1，所述分配模块，包括：

[0208] 分组确定子模块，用于确定位向量表的分组个数m；

[0209] 模式串分配子模块，用于将n个编码后的模式串分配到包含m个分组的位向量表中。

[0210] B14、根据B13所述的装置，所述分组分配子模块，包括：

[0211] 第一计算单元，用于确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

[0212] 第二计算单元，用于根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

[0213] 模式串分配单元，用于将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

[0214] 本发明实施例公开了C15、一种用于数据处理的装置，包括有存储器，以及一个或者一个以上的程序，其中一个或者一个以上程序存储于存储器中，且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：

[0215] 对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合；

[0216] 对所述第一分词集合中的各分词进行编码，得到编码后的模式串；

[0217] 分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现；

[0218] 根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配。

[0219] C16、根据C15所述的装置，所述装置还经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：

[0220] 对所述文本串进行分词，得到所述文本串对应的第二分词集合；

[0221] 对所述第二分词集合中的各分词进行编码，得到编码后的文本串；

[0222] 所述根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，包括：

[0223] 根据所述位向量表，基于位并行算法对所述编码后的模式串与所述编码后的文本串进行匹配。

[0224] C17、根据15所述的装置，所述装置还经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：

[0225] 在所述编码后的模式串的长度大于预设的分组长度时，按照所述分组长度，对所述编码后的模式串进行划分，得到划分后的各子串；

[0226] 在所述各子串中，确定目标子串；

[0227] 所述分配所述编码后的模式串对应的位向量表，包括：

[0228] 分配所述目标子串对应的位向量表；

[0229] 所述根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配，包括：

[0230] 根据所述目标子串对应的位向量表，基于位并行算法对所述目标子串与所述文本串进行匹配，得到匹配串在所述文本串中的位置；

[0231] 根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串。

[0232] C18、根据C17所述的装置，所述在所述各个子串中，确定目标子串，包括：

[0233] 在所述各个子串中，确定词频最小的子串为目标子串。

[0234] C19、根据C17所述的装置，所述根据所述匹配串在所述文本串中的位置，查询所述文本串是否命中所述目标子串对应的模式串，包括：

[0235] 若所述目标子串存在碰撞模式串，则分别计算每个碰撞模式串的词频；

[0236] 根据所述匹配串在所述文本串中的位置，按照碰撞模式串词频从大到小的顺序，

依次查询所述文本串是否命中所述碰撞模式串。

[0237] C20、根据C15所述的装置，若所述待匹配的模式串的个数n大于1，所述分配所述编码后的模式串对应的位向量表，包括：

[0238] 确定位向量表的分组个数m；

[0239] 将n个编码后的模式串分配到包含m个分组的位向量表中。

[0240] C21、根据C20所述的装置，所述将n个编码后的模式串分配到包含m个分组的位向量表中，包括：

[0241] 确定将第i个编码后的模式串分配到第j个分组产生的误判字符串集合；其中，i的取值为1～n，j的取值为1～m；

[0242] 根据所述误判字符串集合中各误判字符串的词频，确定将所述第i个编码后的模式串分配到所述第j个分组产生的第一损失增益；

[0243] 将所述第i个编码后的模式串分配到第一损失增益最小的分组，直到第n个编码后的模式串分配完成。

[0244] 本发明实施例公开了D22、一种机器可读介质，其上存储有指令，当由一个或多个处理器执行时，使得装置执行如A1至A7中一个或多个所述的模式匹配方法。

[0245] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本发明的其它实施方案。本发明旨在涵盖本发明的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本发明的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本发明的真正范围和精神由下面的权利要求指出。

[0246] 应当理解的是，本发明并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

[0247] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

[0248] 以上对本发明所提供的一种模式匹配方法、一种模式匹配装置和一种用于模式匹配的装置，进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

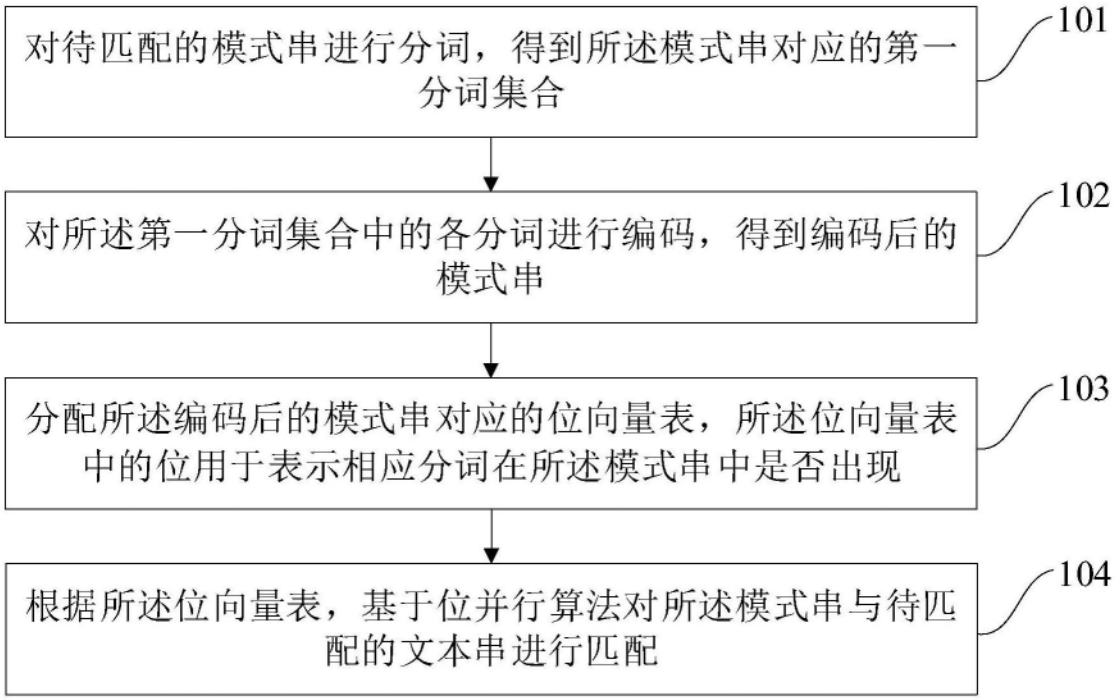

对待匹配的模式串进行分词，得到所述模式串对应的第一分词集合
101
对所述第一分词集合中的各分词进行编码，得到编码后的模式串
102
分配所述编码后的模式串对应的位向量表，所述位向量表中的位用于表示相应分词在所述模式串中是否出现
103
根据所述位向量表，基于位并行算法对所述模式串与待匹配的文本串进行匹配
104

图1

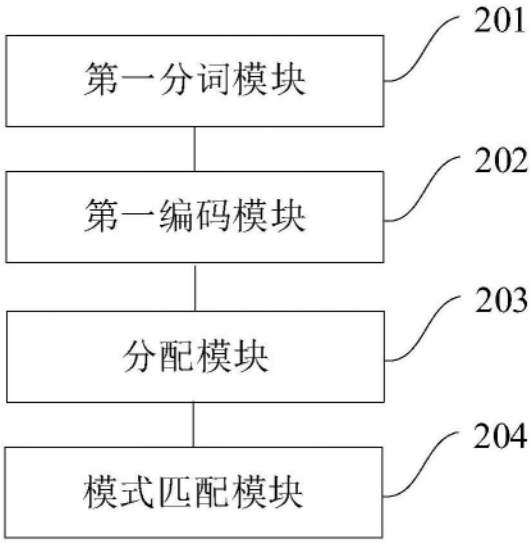

第一分词模块
201
第一编码模块
202
分配模块
203
模式匹配模块
204

图2

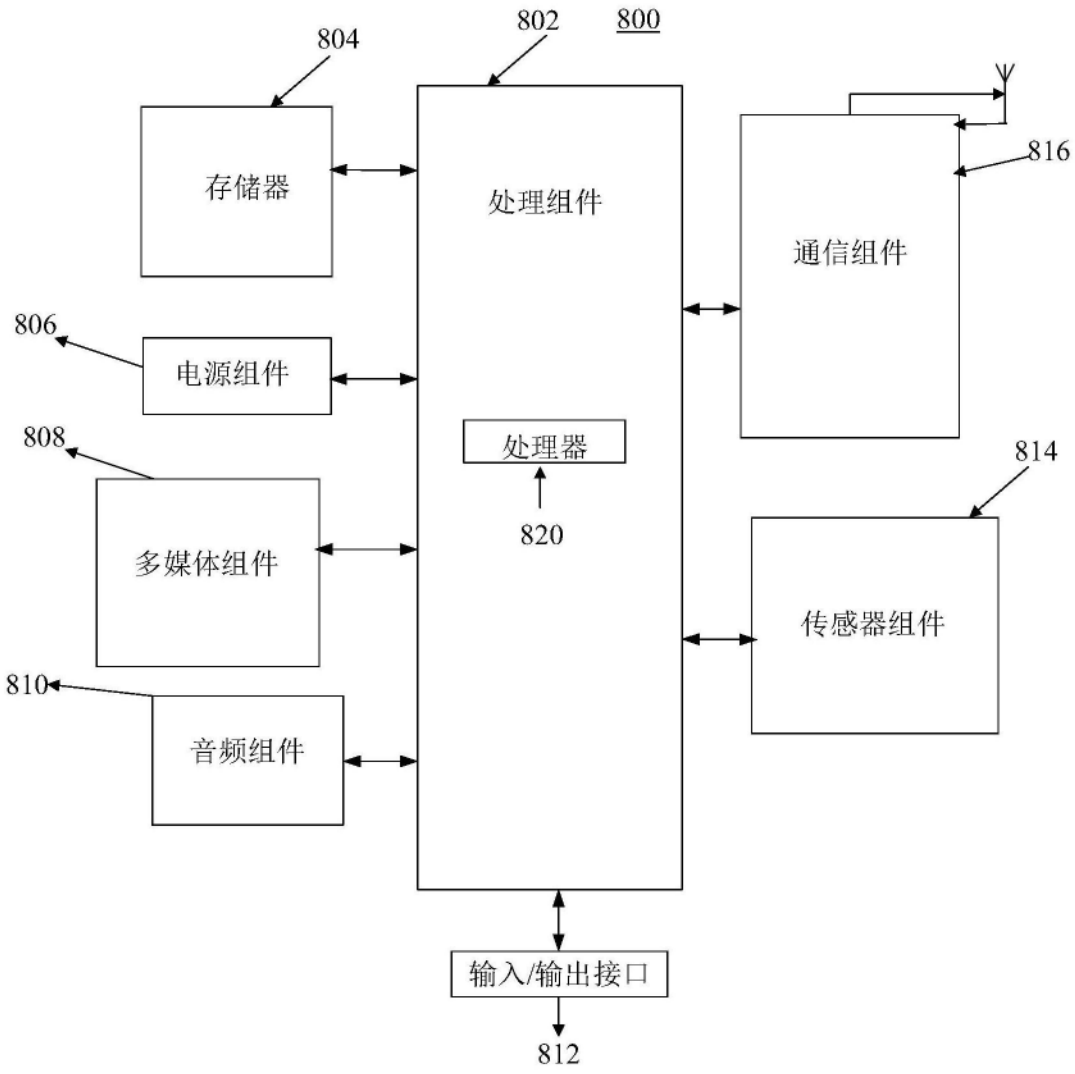

804
802
800
存储器
处理组件
816
通信组件
806
电源组件
处理器
820
808
多媒体组件
814
传感器组件
810
音频组件
输入/输出接口
812

图3

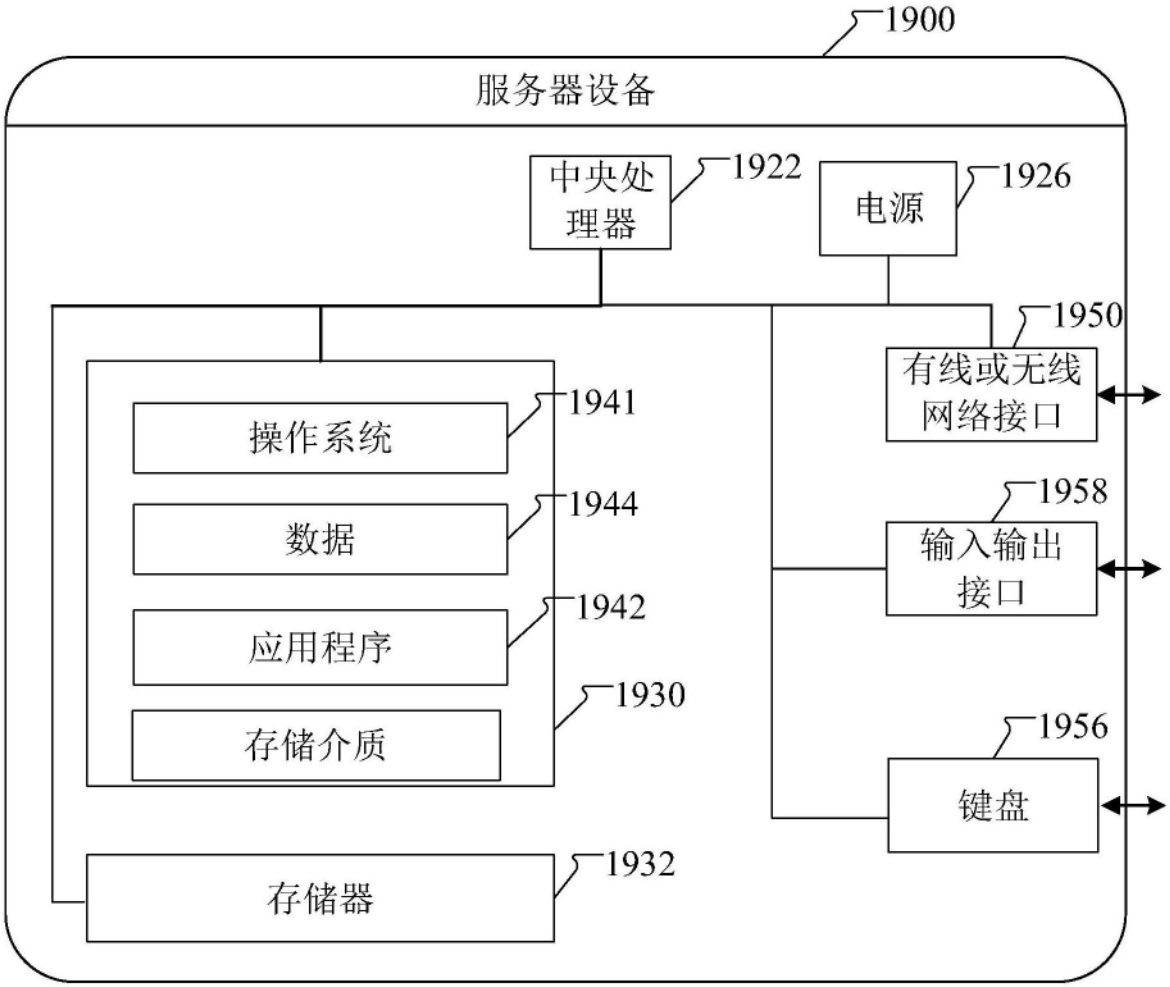

1900
服务器设备
中央处理器
1922
电源
1926
1950
有线或无线网络接口
1941
操作系统
1944
数据
1942
应用程序
1930
存储介质
1958
输入输出接口
1956
键盘
1932
存储器

图4