



República Federativa do Brasil  
Ministério da Economia  
Instituto Nacional da Propriedade Industrial

**(11) PI 0922719-9 B1**



**(22) Data do Depósito:** 17/11/2009

**(45) Data de Concessão:** 19/01/2021

---

**(54) Título:** MÉTODO E DISPOSITIVO PARA INTERCEPTAR CORREIO NÃO SOLICITADO

**(51) Int.Cl.:** H04L 12/58.

**(52) CPC:** H04L 51/12.

**(30) Prioridade Unionista:** 02/12/2008 CN 200810227762.X.

**(73) Titular(es):** TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED.

**(72) Inventor(es):** HUI WANG.

**(86) Pedido PCT:** PCT CN2009074991 de 17/11/2009

**(87) Publicação PCT:** WO 2010/063213 de 10/06/2010

**(85) Data do Início da Fase Nacional:** 02/06/2011

**(57) Resumo:** MÉTODO E DISPOSITIVO PARA INTERCEPTAR CORREIO NÃO SOLICITADO. A presente invenção refere-se a um método e um dispositivo para interceptar um correio não solicitado são fornecidos. O método inclui principalmente: A: obter os dados de texto de um correio que requer processamento de filtração; B: determinar se os dados de texto contêm uma palavra-chave em uma cadeia contida em uma base de dados de carteira para filtração do correio, e se os dados de texto contiverem a palavra-chave na cadeia contida na base de dados de cadeia para filtração de correio, adicionalmente determinar se os dados de texto compreendem uma cadeia que corresponde à palavra-chave contida na base de dados de cadeia; e C: determinar se correio é um correio não solicitado de acordo com um resultado da determinação adicional e de acordo com uma política de determinação predeterminada, e interceptar o correio se o correio for o correio não solicitado. Pelo método e dispositivo, a eficiência de varredura e a velocidade de varredura podem ser melhoradas, e filtração de tempo real para o correio pode ser implementada até mesmo quando a base de cadeia tiver uma dimensão relativamente grande.

Relatório Descritivo da Patente de Invenção para **"MÉTODO E DISPOSITIVO PARA INTERCEPTAR CORREIO NÃO SOLICITADO"**.

CAMPO DA INVENÇÃO

[001] A presente invenção refere-se ao campo de tecnologias de comunicação de rede, e particularmente a um método e dispositivo para interceptar um correio não solicitado (*junk mail*).

ANTECEDENTES DA INVENÇÃO

[002] No campo de correio eletrônico, correios não solicitados difundem-se cada vez mais, que não só aumenta o tempo de processamento de um usuário de correio normal, mas também desperdiça valiosos recursos de um sistema de correio, desse modo obstruindo um processo de obter informação útil por um usuário. Portanto, o problema de correio não solicitado deveria ser solucionado.

[003] No momento, uma técnica de interceptação com base em uma cadeia é tipicamente adotada para impedir o correio não solicitado no sistema de correio. Na técnica de interceptação com base na cadeia, é requerido estabelecer uma base de dados de cadeia. A cadeia na base de dados de cadeia emprega uma palavra simples ou frase existente, e um comprimento da cadeia é relativamente fixado. A base de dados de cadeia necessita ter um certo ciclo de atualização e dimensão, e a dimensão das cadeias passíveis de varredura na base de dados de cadeia frequentemente alcança uma escala de milhões. Em aplicações práticas, usando a cadeia na base de dados de cadeia descrita acima, um correio recebido é filtrado de uma maneira de processamento da varredura sequencial de texto total ou correspondência de expressão regular para determinar se o correio recebido é um correio não solicitado ou um correio normal, e o correio recebido é interceptado se for um correio não solicitado.

[004] Na implementação da presente invenção, o inventor verifica

que existem pelo menos os seguintes problemas na técnica anterior.

[005] Construir a cadeia usando a palavra simples ou frase existente pode levar a uma taxa de falso positivo relativamente séria porque tal palavra simples ou frase existente não só é apresentada no correio não solicitado, mas também às vezes no correio normal, desse modo levando à determinação falsa.

[006] Uma vez que uma cadeia completa na base de dados de cadeia é usada para filtrar o correio, a maneira de processamento acima descrita de varredura sequencial de texto total ou equiparação de expressão regular é ineficiente quando a dimensão da base de dados de cadeia for relativamente grande, e filtração de tempo real para o correio recebido não pode ser implementada que significativamente afeta a experiência de uso do usuário.

#### SUMÁRIO DA INVENÇÃO

[007] Exemplos da presente invenção fornecem um método e dispositivo para interceptar um correio não solicitado, para diminuir uma taxa de falso positivo do correio não solicitado e melhorar uma eficiência de filtração do correio.

[008] Um método para interceptar correio não solicitado, que inclui:

A: obter dados de texto de um correio que requer processamento de filtração;

B: determinar se os dados de texto contêm uma palavra-chave em uma cadeia contida em uma base de dados de cadeia para filtração de correio, e se os dados de texto contiverem a palavra-chave na cadeia contida na base de dados de cadeia para filtração de correio, adicionalmente determinar se os dados de texto contêm uma cadeia correspondendo à palavra-chave contida na base de dados de cadeia;

e

C: determinar se o correio é um correio não solicitado de acordo com um resultado da determinação adicional e de acordo com

uma política de determinação predeterminada, e interceptar o correio se o correio for o correio não solicitado.

[009] Um dispositivo para interceptar correio não solicitado inclui:

[0010] um módulo de obtenção de dados de texto, configurado para obter dados de texto de um correio que requer processamento de filtração;

[0011] um módulo de determinação de caractere, configurado para determinar se os dados de texto contêm uma palavra-chave em uma cadeia contida em uma base de dados de cadeia para filtração de correio, e se os dados de texto contêm a palavra-chave na cadeia contida na base de dados de cadeia para filtração de correio, adicionalmente determinar se os dados de texto contêm uma cadeia que corresponde à palavra-chave contida na base de dados de cadeia; e

[0012] um módulo de processamento de correio, configurado para determinar, de acordo com um resultado da determinação adicional do módulo de determinação de caractere como também uma política de determinação predeterminada, se o correio é o correio não solicitado, e interceptar o correio se o correio for o correio não solicitado.

[0013] Pode ser visto das soluções técnicas acima fornecidas pelos exemplos da presente invenção que nos exemplos da presente invenção, os dados de texto do correio são varridos de acordo com a palavra-chave, os dados de texto do correio são depois varridos de acordo com a cadeia que corresponde à palavra-chave após equiparação da palavra-chave, desse modo uma velocidade e eficiência da varredura podem ser melhoradas, e filtração de tempo real para o correio pode ser implementada até mesmo quando a base de dados de cadeia tiver uma dimensão relativamente grande.

#### BREVE DESCRIÇÃO DOS DESENHOS

[0014] Para mais claramente explicar as soluções técnicas nos exemplos da presente invenção, os desenhos em anexo requeridos na

descrição dos exemplos são concisamente listados abaixo. É evidente que os desenhos em anexo na descrição abaixo são meramente alguns exemplos da presente invenção, e para aqueles normalmente versados na técnica, outros desenhos em anexo podem também ser obtidos de acordo com estes desenhos em anexo sem exercer qualquer etapa inventiva. Em que,

[0015] figura 1 é um fluxograma que ilustra um método para interceptar um correio não solicitado em um exemplo da presente invenção; e

[0016] figura 2 é um diagrama estrutural que ilustra a implementação específica de um dispositivo para interceptar um correio não solicitado em outro exemplo da presente invenção.

#### DESCRIÇÃO DETALHADA DA INVENÇÃO

[0017] Nos exemplos da presente invenção, os dados de texto de um correio que requer processamento de filtração são obtidos; é determinado se os dados de texto obtidos do correio contêm uma palavra-chave em uma cadeia em uma base de dados de cadeia para filtração de correio, e quando os dados de texto obtidos contiverem a palavra-chave, é ainda determinado se os dados de texto contêm a cadeia que corresponde à palavra-chave na base de dados de cadeia. De acordo com um resultado de determinação com relação se os dados de texto contêm a cadeia que corresponde à palavra-chave na base de dados de cadeia e de acordo com uma política de determinação predeterminada, é determinado se o correio é um correio não solicitado, e o correio é interceptado se o correio for o correio não solicitado.

[0018] Ainda, após o correio que requer o processamento de filtração ser recebido, um título e os conteúdos do corpo principal do correio são obtidos; em seguida o título e os conteúdos do corpo principal são alinhavados para obter um trecho dos dados de texto; e os dados de texto obtidos são determinados como os dados de texto do

correio que requer o processamento de filtração. Preferivelmente, os dados de texto podem ser armazenados.

[0019] Ainda, a cadeia contida na base de dados de cadeia é construída por uma ou mais unidades de caractere. Uma unidade de caractere inclui pelo menos um de uma palavra inglesa, uma palavra simples chinesa, uma letra inglesa simples, uma metade da palavra simples chinesa ou uma pontuação de largura total/meia-largura.

[0020] Ainda, a base de dados de cadeia corresponde a uma tabela principal de valor *hash* e uma tabela de ligação de valor *hash* onde a palavra-chave na cadeia contida na base de dados de cadeia e a informação de comprimento da cadeia que corresponde à palavra-chave são armazenadas na tabela principal de valor *hash*, e a informação de construção de caractere completa da cadeia que corresponde à palavra-chave é armazenada na tabela de ligação de valor *hash*.

[0021] Quando uma operação de determinação descrita acima for executada, o detalhe é: extrair um número prefixado de caracteres a partir de uma primeira unidade de caractere dos dados de texto, detectar se a tabela principal de valor *hash* contém a palavra-chave que é igual ao número prefixado de caracteres, e se sim, obter a informação de comprimento (especificamente, um valor de comprimento) correspondendo à palavra-chave, tirar a cadeia correspondente dos dados de texto de acordo com a informação de comprimento, detectar se a tabela de ligação de valor *hash* contém a cadeia tirada, e se sim, determinar que os dados de texto são atingidos por varredura por uma vez, e registrar o número de vezes que os dados de texto são atingidos por varredura, como também informação da palavra-chave correspondente e cadeia.

[0022] Se a tabela principal de valor *hash* não contiver a palavra-chave que é igual ao número prefixado de caracteres, ou se a tabela de

ligação de valor *hash* não contiver a cadeia tirada, o número prefixado de caracteres é tirado após deslocar para trás por uma unidade de caractere da primeira unidade de caractere dos dados de texto, e os caracteres tirados de acordo com uma operação de processamento são processados para o número prefixado de caracteres tirados da primeira unidade de caractere dos dados de texto até o último número prefixado de caracteres nos dados de texto serem detectados.

[0023] Ainda, a tabela principal de valor *hash* e a tabela de ligação de valor *hash* são estabelecidas por: tirar o número prefixado de caracteres a partir do primeiro caractere em uma primeira cadeia contida na base de dados de cadeia, tirar os caracteres tirados como uma palavra-chave, determinar se o número prefixado de caracteres da primeira unidade de caractere em outra cadeia diferente da primeira cadeia na base de dados de cadeia é igual à palavra-chave, e se a mesma, registrar informação de comprimento da outra cadeia e da palavra-chave na tabela principal de valor *hash* e registrar a informação de construção de caractere completa da outra cadeia na tabela de ligação de valor *hash*; e depois

[0024] adicionalmente determinar uma segunda cadeia diferente de uma cadeia registrada na tabela de ligação de valor *hash* na base de dados de cadeia, e processar a segunda cadeia de acordo com uma operação de processamento para o número prefixado de caracteres tirados da primeira cadeia, até registrar todas as seções de caracteres tirados a partir das respectivas primeiras unidades de caractere de todas as cadeias na base de dados de cadeia e informação de comprimento das mesmas na tabela principal de valor *hash*, e registrar a respectiva informação de construção de caractere completa de todas as cadeias correspondentes na tabela de ligação de valor *hash*.

[0025] Ainda, a determinação se o correio é um correio não solicitado inclui: obter o número registrado de vezes que os dados de

texto são atingidos por varredura, como também a informação registrada sobre a palavra-chave correspondente e a cadeia é registrada quando os dados de texto contiverem a cadeia que corresponde à palavra-chave na base de dados de cadeia, e é depois obtida; e

[0026] de acordo com o número registrado de vezes que os dados de texto são atingidos por varredura como também a informação registrada sobre a palavra-chave correspondente e a cadeia, é determinado se o correio é o correio não solicitado com base na política de determinação predeterminada, e o correio é interceptado se o correio for o correio não solicitado.

[0027] Ainda, a política de determinação predeterminada contém: o correio é determinado como o correio não solicitado quando o número de vezes que os dados de texto são atingidos por varredura for maior que um número prefixado de vezes; ou se a informação da cadeia for o comprimento da cadeia atingido por varredura, a política de determinação predeterminada inclui: o correio é determinado como o correio não solicitado quando o número de vezes que os dados de texto são atingidos por varredura for maior que o número prefixado de vezes e o comprimento da cadeia atingido por varredura for maior que um comprimento prefixado.

[0028] Para facilitar a compreensão dos exemplos da presente invenção, uma explanação adicional é feita doravante através de vários exemplos específicos em combinação com os desenhos em anexo, e os respectivos exemplos não são intencionados a limitar os exemplos da presente invenção.

[0029] Um esquema de valor *hash* é uma estrutura de armazenamento. No esquema de valor *hash*, uma relação correspondente é estabelecida entre uma posição de armazenamento de dados e a palavra-chave dos dados, e um conjunto das palavras-

chaves é mapeado para um conjunto de localização através da relação correspondente. Ajuste da relação correspondente é flexível, contanto que o tamanho do conjunto de localização não ultrapasse uma faixa permissível. O esquema de valor *hash* tipicamente inclui uma tabela principal de valor *hash* e uma tabela de ligação de valor *hash*. Em aplicações práticas, é requerido constituir a tabela principal de valor *hash* e a tabela de ligação de valor *hash* de acordo com uma situação atual.

[0030] De acordo com um exemplo, um procedimento de processamento de um método para interceptar um correio não solicitado é mostrado na Figura 1, e o método inclui as etapas de processamento como segue:

[0031] Etapa 11: Os dados de texto do correio que requer o processamento de filtração são obtidos.

[0032] O detalhe é: após o correio que requer o processamento de filtração ser recebido, decodificar o correio e obter o título e o conteúdo do corpo principal do correio; obter um trecho dos dados de texto alinhando o título e o conteúdo do corpo principal diretamente; e determinar os dados de texto obtidos como os dados de texto do correio que requer o processamento de filtração na Etapa 11.

[0033] Aqui para facilitar a interceptação na etapa seguinte que é especificamente mostrada na Etapa 13 abaixo os dados de texto podem ser primeiro temporariamente armazenados.

[0034] Etapa 12: De acordo com uma base de dados de cadeia carregada, a tabela principal de valor *hash* e a tabela de ligação de valor *hash* são estabelecidas.

[0035] Aqui, uma vez que a tabela principal de valor *hash* e a tabela de ligação de valor *hash* são estabelecidas de acordo com a base de dados de cadeia, pode ser considerado que a base de dados de cadeia tem uma relação correspondente à tabela principal de valor *hash* e à

tabela de ligação de valor *hash*.

[0036] Deveria ser explicado que a cadeia contida na base de dados de cadeia é construída por uma ou mais unidades de caractere. Especificamente, a unidade de caractere pode ser pelo menos um de uma palavra inglesa, uma palavra simples chinesa, uma carta inglesa simples, uma metade da palavra simples chinesa ou uma pontuação de largura total/meia-largura. Pode ser visto que a cadeia contida na base de dados de cadeia pode não ser uma palavra simples ou frase existente, mas uma seção de cadeia tendo uma estrutura flexível. A seção de cadeia pode ser pelo menos uma ou qualquer combinação da palavra inglesa, da palavra simples chinesa e da pontuação. Tipicamente, em aplicações práticas, a cadeia existe principalmente em um correio não solicitado ou um correio normal. Preferivelmente, uma situação que a cadeia contida na base de dados de cadeia é apresentada no correio não solicitado é tomada como um exemplo. Deveria ser observado que este exemplo leva a situação que a cadeia contida na base de dados de cadeia é apresentada no correio não solicitado como um exemplo. Em atenção a um escopo de aplicação dos exemplos da presente invenção, a cadeia contida na base de dados de cadeia descrita acima pode também exemplificadamente ser a cadeia que existe no correio normal, isto é, as cadeias no correio normal e no correio não solicitado são simultaneamente usadas. Preferivelmente, quando ambas forem simultaneamente usadas, os dados de texto específicos podem ser varridos e determinados usando um método tal como qualquer algoritmo de classificação estatístico e/ou algoritmo de classificação de inteligência artificial. Por exemplo, os dois tipos de cadeias no correio normal e no correio não solicitado podem ser treinados e testados usando um algoritmo Bayesiano para obter um modelo de classificação, e o modelo de classificação é usado para executar a determinação subsequente do conteúdo de texto de um

correio. Portanto, pode ser visto que a Figura 1 meramente mostra um exemplo, e não é intencionado a limitar a aplicação dos exemplos da presente invenção.

[0037] No exemplo, o esquema de valor *hash* descrito acima é introduzido, e de acordo com a base de dados de cadeia carregada, a tabela principal de valor *hash* e a tabela de ligação de valor *hash* são estabelecidas. Um processo para estabelecer a tabela principal de valor *hash* e a tabela de ligação de valor *hash* é como segue:

[0038] as cadeias na base de dados de cadeia descrita acima são varridas sequencialmente desde o princípio da base de dados de cadeia. Primeiramente, os primeiros *n* caracteres de uma primeira cadeia são tomados como um índice de valor *hash* de primeiro nível. Para conveniência de descrição, é suposto que *n* seja 2. O índice de valor *hash* de primeiro nível é depois determinado como a palavra-chave. Por exemplo, a palavra-chave é "SanLu" que representa uma palavra chinesa formada por dois caracteres chineses. Depois, com a palavra-chave como um índice, outra cadeia diferente da primeira cadeia na base de dados de cadeia descrita acima é pesquisada, e se os 2 primeiros caracteres da outra cadeia são iguais à palavra-chave é determinado. Se os 2 primeiros caracteres da outra cadeia forem iguais à palavra-chave, a informação de construção de caractere completa e a informação de comprimento da cadeia de antera são obtidas.

[0039] Preferivelmente, neste exemplo, a informação de comprimento de todas as cadeias que levam a palavra-chave, por exemplo "SanLu", como os 2 primeiros caracteres chineses pode ser armazenada na tabela principal de valor *hash*. Uma estrutura da tabela principal de valor *hash* é como mostrada na Tabela 1 listada abaixo. Depois disso, a respectiva informação de construção de caractere completa de todas as cadeias que levam a palavra-chave, por exemplo, "SanLu", como os 2 primeiros caracteres é armazenada na tabela de

ligação de valor *hash*. Uma estrutura da tabela de ligação de valor *hash* é como mostrada na Tabela 2 listada abaixo. Portanto, pode ser visto que aquela palavra-chave corresponde a uma tabela de ligação de valor *hash*. No esquema de valor *hash*, há apenas uma tabela principal de valor *hash* em que todas palavras-chaves e a informação de comprimento das cadeias que leva cada palavra-chave como os primeiros n caracteres são armazenadas. Pode haver tabelas de ligação de valor *hash* múltiplas, que correspondem às respectivas palavras-chaves na tabela principal de valor *hash*.

**Tabela 1:** Tabela principal de valor *hash*

Palavra-chave	Valor de comprimento			
SanLu	4	5	6	...
...				

**Tabela 2:** Tabela de ligação de valor *hash*

Leite SanLu
Leite puro SanLu
Leite infantil SanLu
...

[0040] Após o processamento acima tal como tirar a palavra-chave para a primeira cadeia e encher a Tabela 1 e Tabela 2 de acordo com a palavra-chave, o processamento acima tal como tirar a palavra-chave e encher a Tabela 1 e Tabela 2 de acordo com a palavra-chave é depois executado para outra cadeia diferente das cadeias registradas na tabela de ligação de valor *hash* mostrada na Tabela 2 na base de dados de cadeia descrita acima, até a informação de comprimento e os primeiros n caracteres de todas as cadeias na base de dados de cadeia serem registrados na tabela principal de valor *hash* e a respectiva informação de construção de caractere completa de todas as cadeias ser armazenada na tabela de ligação de valor *hash*.

[0041] Desse modo, através das etapas descritas acima, a tabela principal de valor *hash* e as tabelas de ligação de valor *hash* correspondentes pode ser estabelecida com respeito à base de dados

de cadeia.

[0042] Etapa 13: Os dados de texto do correio são varridos usando a tabela principal de valor *hash* e a tabela de ligação de valor *hash*, se o correio é o correio não solicitado é determinado de acordo com um resultado de varredura e uma política de determinação predeterminada, e o correio é interceptado se o correio for o correio não solicitado.

[0043] Após a tabela principal de valor *hash* e a tabela de ligação de valor *hash* descritas acima serem estabelecidas, para os dados de texto do correio que requer o processamento de filtração, uma cadeia construída pelos primeiros *n* caracteres (onde *n* pode especificamente ser 2 ou outro valor) é tirada a partir do primeiro caractere dos dados de texto e é detectado se uma palavra-chave que é igual à cadeia tirada existe na tabela principal de valor *hash* estabelecida. Se tal palavra-chave existir, um primeiro valor de comprimento correspondendo à cadeia é obtido. Depois, a cadeia correspondente é tirada dos dados de texto de acordo com o primeiro valor de comprimento, e é detectado se a cadeia tirada existe na tabela de ligação de valor *hash*. Se tal cadeia existir, é determinado que a varredura atinge os dados de texto uma vez e a informação tal como a palavra-chave correspondente e a cadeia atingida pela varredura é registrada; se tal cadeia não existir, nenhuma informação será registrada. A tabela principal de valor *hash* é verificada novamente para um próximo valor de comprimento que corresponde à cadeia, até todos os valores de comprimento que correspondem à cadeia serem detectados.

[0044] Se a palavra-chave que é igual à cadeia tirada não existir na tabela principal de valor *hash*, a tabela de ligação de valor *hash* não necessita ser verificada. Depois, a partir do segundo caractere dos dados de texto, a cadeia com 2 caracteres é tirada. E é detectado se a tabela principal de valor *hash* inclui uma palavra-chave que é igual à cadeia tirada a partir do segundo caractere dos dados de texto, e o

processo de detecção e determinação acima com respeito à cadeia tirada a partir do primeiro caractere é atingido até a cadeia construída pelos 2 últimos caracteres dos dados de texto ser detectada.

[0045] Depois, de acordo com a informação registrada sobre o número de vezes que a varredura atinge os dados de texto e a informação tal como a palavra-chave correspondente e a cadeia atingida pela varredura, se o correio é o correio não solicitado é determinado com base na política de determinação predeterminada. A política de determinação predeterminada é projetada de acordo com a situação atual, e a política de determinação pode ser como segue: se o número de vezes que os dados de texto são atingidos pela varredura for maior que 5, o correio é determinado como o correio não solicitado, ou se o número de vezes que os dados de texto são atingidos a varredura é maior que 4 e o comprimento da cadeia atingida pela varredura é maior que 4, o correio é determinado como o correio não solicitado.

[0046] A política de determinação predeterminada deveria assegurar que uma taxa de falso positivo inteira deveria ser menor que um índice de taxa de falso positivo aceitável, por exemplo 0,1%, e uma taxa de interceptação inteira deveria ser maior que um índice de taxa de interceptação aceitável, por exemplo 70%.

[0047] Depois, o correio não solicitado determinado é interceptado, e o correio normal que não é o correio não solicitado passa.

[0048] No processo acima para varrer o correio, os dados de texto do correio são primeiro varridos de acordo com a palavra-chave, e após ser verificado que os dados de texto do correio contêm a palavra-chave, os dados de texto do correio são depois varridos de acordo com a cadeia que corresponde à palavra-chave. Desse modo, uma velocidade e eficiência da varredura podem ser melhoradas.

[0049] Outro exemplo da presente invenção também provê um dispositivo para interceptar um correio não solicitado. Sua estrutura de

implementação específica é como mostrada na figura 2. O dispositivo pode especificamente incluir o seguinte:

[0050] um módulo de obtenção de dados de texto 21, configurado para obter dados de texto de um correio que requer processamento de filtração;

[0051] um módulo de determinação de caractere 22, configurado para determinar se os dados de texto contêm uma palavra-chave em uma cadeia contida em uma base de dados de cadeia para filtração de correio, e se sim, adicionalmente determinar se os dados de texto contêm a cadeia que corresponde à palavra-chave contida na base de dados de cadeia; e

[0052] um módulo de processamento de correio 23, configurado: de acordo com um resultado de determinação adicional do módulo de determinação de caractere 22 e uma política de determinação predeterminada, determinar se o correio é um correio não solicitado, e interceptar o correio se for o correio não solicitado. Aqui, o resultado de determinação adicional do módulo de determinação de caractere 22 pode ser especificamente um resultado de determinação que considera se os dados de texto contêm a cadeia que corresponde à palavra-chave contida na base de dados de cadeia.

[0053] O módulo de determinação de caractere 22 pode especificamente incluir:

[0054] um módulo de estabelecimento de tabela de valor *hash* 221, configurado para estabelecer uma tabela principal de valor *hash* e uma tabela de ligação de valor *hash* que correspondem à base de dados de cadeia, em que a tabela principal de valor *hash* armazena a palavra-chave na cadeia contida na base de dados de cadeia e a informação de comprimento da cadeia que corresponde à palavra-chave, e a tabela de ligação de valor *hash* armazena a informação de construção de caractere completa da cadeia que corresponde à palavra-chave; e

[0055] um módulo de processamento de varredura 222, configurado para extrair um número prefixado de caracteres a partir de uma primeira unidade de caractere dos dados de texto, detectar se a tabela principal de valor *hash* contém a palavra-chave que é igual ao número prefixado de caracteres, e se sim, obter a informação de comprimento (especificamente, um valor de comprimento) correspondendo à palavra-chave, tirar a cadeia correspondente dos dados de texto de acordo com a informação de comprimento, detectar se a cadeia tirada existe na tabela de ligação de valor *hash*, e se sim, determinar que os dados de texto são atingidos pela varredura por uma vez, e registrar o número de vezes que os dados de texto são atingidos pela varredura como também informação da palavra-chave e cadeia correspondentes.

[0056] Se a tabela principal de valor *hash* não contiver a palavra-chave que é igual ao número prefixado de caracteres, ou se a tabela de ligação de valor *hash* não contiver a cadeia tirada, o número prefixado de caracteres é tirado dos dados de texto após deslocar para trás por uma unidade de caractere do primeiro caractere dos dados de texto, e os caracteres tirados após deslocar para trás por uma unidade de caractere do primeiro caractere dos dados de texto de acordo com uma operação de processamento para o número prefixado de caracteres tirados do primeiro caractere dos dados de texto, até o último número prefixado de caracteres nos dados de texto, são detectados.

[0057] O módulo de processamento de correio 23 especificamente inclui:

[0058] Um módulo de obtenção de informação de varredura 231, configurado para obter a informação registrada sobre o número de vezes que os dados de texto são atingidos por varredura, como também a informação registrada sobre a palavra-chave e cadeia correspondentes. Especificamente, a informação sobre o número de vezes que os dados de texto são atingidos por varredura, como também

a informação sobre a palavra-chave e cadeia correspondentes são registradas quando os dados de texto contiverem a cadeia que corresponde à palavra-chave na base de dados de cadeia; e

[0059] um módulo de determinação e interceptação 232, configurado para determinar, de acordo com a informação sobre o número de vezes que os dados de texto são atingidos por varredura como também de acordo com a informação da palavra-chave e cadeia correspondentes, se o correio é o correio não solicitado com base na política de determinação predeterminada; e interceptar o correio se o correio for determinado como o correio não solicitado.

[0060] Aqueles normalmente versados na técnica podem entender que todo ou parte do procedimento no método nos exemplos descritos acima pode ser implementado por um programa de computação instruindo hardware relevante. O programa pode ser armazenado em um meio de armazenamento legível por computador. Quando o programa for executado, o procedimento nos exemplos para os respectivos métodos descritos acima pode ser implementado. Especificamente, o meio de armazenamento pode ser um disco magnético, um disco óptico, uma Memória Exclusiva de Leitura (ROM) ou uma Memória de Acesso Aleatório (RAM), etc.

[0061] Para resumir, usando a seção de cadeia tendo a estrutura flexível que é apresentada apenas no correio não solicitado em vez de usar uma palavra simples ou frase, os exemplos da presente invenção podem solucionar o falso problema de determinação na técnica anterior, e tem uma taxa relativamente baixa de falso positivo e uma taxa de interceptação relativamente alta.

[0062] Usando a tabela principal de valor *hash* e a tabela de ligação de valor *hash* no esquema de valor *hash*, os exemplos da presente invenção varrem os dados de texto do correio que pode grandemente melhorar a eficiência da varredura e melhora a velocidade da varredura,

e pode implementar filtração de tempo real até mesmo para o correio quando a base de dados de cadeia tiver uma dimensão relativamente grande.

[0063] O antecedente são meramente exemplos preferidos da presente invenção, e o escopo da presente invenção não é limitado a estes. Quaisquer variações ou alterações facilmente feitas sem abandono do escopo técnico da presente invenção por aqueles versados na técnica devem ser abrangidas dentro do escopo da presente invenção. Portanto, o escopo da presente invenção deve ser como definido pelas concretizações em anexo.

## REIVINDICAÇÕES

1. Método para interceptar um correio não solicitado, **caracterizado pelo fato de que** compreende etapas de:

A: obter dados de texto de um correio que requer processamento de filtração (11);

B: determinar se os dados de texto compreendem uma palavra-chave em uma cadeia contida em uma base de dados de cadeia para filtração de correio, e se os dados de texto compreenderem a palavra-chave na cadeia contida na base de dados de cadeia para filtração de correio, adicionalmente determinar se os dados de texto compreendem uma cadeia que corresponde à palavra-chave contida na base de dados de cadeia; e

C: determinar se o correio é um correio não solicitado de acordo com um resultado da determinação adicional e de acordo com uma política de determinação predeterminada, e interceptar o correio se o correio for o correio não solicitado (13);

em que a base de dados de cadeia corresponde a uma tabela principal de valor hash e uma tabela de ligação de valor hash;

em que a tabela principal de valor hash armazena a palavra-chave na cadeia contida na base de dados de cadeia e informações de comprimento da cadeia correspondendo à palavra-chave, e a tabela de ligação de valor hash armazena informações de construção de caractere completas da cadeia correspondendo à palavra-chave;

em que a etapa B compreende:

B1: extrair um número prefixado de caracteres a partir de um primeiro caractere dos dados de texto, detectar se a tabela principal de valor *hash* contém uma palavra-chave que é igual ao número prefixado de caracteres, e se a tabela principal de valor *hash* contiver uma palavra-chave que é igual ao número prefixado de caracteres, obter a informação de comprimento que corresponde à palavra-chave, tirar uma

cadeia dos dados de texto de acordo com a informação de comprimento, detectar se a tabela de ligação de valor *hash* contém a cadeia tirada; e se a tabela de ligação de valor *hash* contiver a cadeia tirada, determinar que os dados de texto são atingidos por varredura por uma vez, e registrar o número de vezes que os dados de texto são atingidos por varredura assim como informação sobre a palavra-chave e a cadeia que corresponde à palavra-chave; e

B2: se a tabela principal de valor *hash* não contiver a palavra-chave que é igual ao número prefixado de caracteres, ou se a tabela de ligação de valor *hash* não contiver a cadeia tirada, tirar o número prefixado de caracteres após deslocar para trás por uma unidade de caractere do primeiro caractere dos dados de texto, e processar os caracteres tirados de acordo com uma operação de processamento para o número prefixado de caracteres tirados do primeiro caractere dos dados de texto na Etapa B1, até detectar um último número prefixado de caracteres nos dados de texto.

2. Método, de acordo com a reivindicação 1, **caracterizado pelo fato de que** a Etapa A compreende:

após receber o correio que requer o processamento de filtração, obter um título e conteúdo do corpo principal do correio; alinhar o título e o conteúdo do corpo principal para obter dados de texto; e determinar os dados de texto obtidos como os dados de texto do correio que requerem o processamento de filtração.

3. Método, de acordo com a reivindicação 1, **caracterizado pelo fato de que** a cadeia contida na base de dados de cadeia é construída por uma ou mais unidades de caractere; em que a unidade de caractere compreende pelo menos um de uma palavra inglesa, uma palavra simples chinesa, uma letra inglesa simples, uma metade da palavra simples chinesa ou uma pontuação de largura total/meia-largura.

4. Método, de acordo com a reivindicação 1, **caracterizado pelo fato de que** a tabela principal de valor *hash* e a tabela de ligação de valor *hash* são estabelecidas por:

B01: tirar o número prefixado de caracteres a partir da primeira unidade de caractere em uma primeira cadeia contida na base de dados de cadeia, considerar os caracteres tirados como a palavra-chave, determinar se o número prefixado de caracteres da primeira unidade de caractere em outra cadeia diferente da primeira cadeia na base de dados de cadeia é igual à palavra-chave, e se igual, registrar a palavra-chave e a informação de comprimento da outra cadeia na tabela principal de valor *hash* e registrar a informação de construção de caractere completa da outra cadeia na tabela de ligação de valor *hash*;  
e

B02: adicionalmente determinar uma segunda cadeia diferente de uma cadeia registrada na tabela de ligação de valor *hash* na base de dados de cadeia, e processar a segunda cadeia de acordo com uma operação de processamento para a primeira cadeia na Etapa B01, até terminar a operação de processamento para a primeira cadeia na Etapa B01 para todas as cadeias contida na base de dados de cadeia.

5. Método, de acordo com a reivindicação 1, **caracterizado pelo fato de que** a Etapa C compreende:

C1: obter o número registrado de vezes que os dados de texto são atingidos por varredura, assim como a informação registrada sobre a palavra-chave e a cadeia que corresponde à palavra-chave; e

C2: de acordo com o número registrado de vezes que os dados de texto são atingidos por varredura assim como a informação registrada sobre a palavra-chave e a cadeia que corresponde à palavra-chave, determinar se o correio é o correio não solicitado com base na política de determinação predeterminada, e interceptar o correio se o

correio for o correio não solicitado.

6. Método, de acordo com a reivindicação 5, **caracterizado pelo fato de que** a política de determinação predeterminada compreende: o correio é determinado como o correio não solicitado quando o número de vezes que os dados de texto são atingidos por varredura for maior que um número prefixado de vezes; ou

se a informação sobre a cadeia na Etapa C1 for o comprimento da cadeia atingida por varredura, a política de determinação predeterminada na Etapa C2 compreende: o correio é determinado como o correio não solicitado quando o número de vezes que os dados de texto são atingidos por varredura for maior que o número prefixado de vezes e o comprimento das cadeias atingida por varredura for maior que um comprimento prefixado.

7. Dispositivo para interceptar um correio não solicitado, **caracterizado pelo fato que** de compreende:

um módulo de obtenção de dados de texto (21), configurado para obter dados de texto de um correio que requer processamento de filtração;

um módulo de determinação de caractere (22), configurado para determinar se os dados de texto compreendem uma palavra-chave em uma cadeia contida em uma base de dados de cadeia para filtração de correio, e se os dados de texto compreenderem a palavra-chave na cadeia contida na base de dados de cadeia para filtração de correio, adicionalmente determinar se os dados de texto compreendem uma cadeia que corresponde à palavra-chave contida na base de dados de cadeia; e

um módulo de processamento de correio (23), configurado para determinar, de acordo com um resultado da determinação adicional do módulo de determinação de caractere (22) assim como uma política de determinação predeterminada, se o correio é o correio não solicitado,

e interceptar o correio se o correio for o correio não solicitado;

em que o módulo de determinação de caractere (22) compreende:

um módulo de estabelecimento da tabela de valor *hash* (221), configurado para estabelecer uma tabela principal de valor *hash* e uma tabela de ligação de valor *hash* que correspondem à base de dados de cadeia, em que a tabela principal de valor *hash* armazena a palavra-chave na cadeia contida na base de dados de cadeia e a informação de comprimento da cadeia que corresponde à palavra-chave, e a tabela de ligação de valor *hash* armazena a informação de construção de caractere completa da cadeia que corresponde à palavra-chave; e

um módulo de processamento de varredura (222), configurado para extrair um número prefixado de caracteres a partir de uma primeira unidade de caractere dos dados de texto, detectar se a tabela principal de valor *hash* contém a palavra-chave que é igual ao número prefixado de caracteres, e se a tabela principal de valor *hash* contiver uma palavra-chave que é igual ao número prefixado de caracteres, obter a informação de comprimento que corresponde à palavra-chave, tirar uma cadeia dos dados de texto de acordo com a informação de comprimento, detectar se a tabela de ligação de valor *hash* contém a cadeia tirada, e se a tabela de ligação de valor *hash* contiver a cadeia tirada, determinar que os dados de texto são atingidos por varredura por uma vez, e registrar o número de vezes que os dados de texto são atingidos por varredura assim como a informação sobre a palavra-chave e a cadeia que corresponde à palavra-chave; e se a tabela principal de valor *hash* não contiver a palavra-chave que é igual ao número prefixado de caracteres ou se a tabela de ligação de valor *hash* não contiver a cadeia tirada, configurado para tirar o número prefixado de caracteres após deslocar para trás por uma unidade de

caractere do primeiro caractere dos dados de texto, e processar os caracteres tirados após deslocar para trás por uma unidade de caractere a partir do primeiro caractere dos dados de texto de acordo com uma operação de processamento para o número prefixado de caracteres tirados a partir da primeira unidade de caractere dos dados de texto até detectar um último número prefixado de caracteres nos dados de texto.

8. Dispositivo, de acordo com a reivindicação 7, **caracterizado pelo fato de que** o módulo de processamento de correio (23) compreende:

um módulo de obtenção de informação de varredura (231), configurado para obter o número registrado de vezes que os dados de texto são atingidos por varredura, assim como a informação registrada sobre a palavra-chave e a cadeia que corresponde à palavra-chave; e

um módulo de determinação e interceptação (232), configurado para determinar, de acordo com o número registrado de vezes que os dados de texto são atingidos por varredura assim como de acordo com a informação registrada sobre a palavra-chave e a cadeia que corresponde à palavra-chave, se o correio é o correio não solicitado com base na política de determinação predeterminada, e interceptar o correio se o correio for o correio não solicitado.

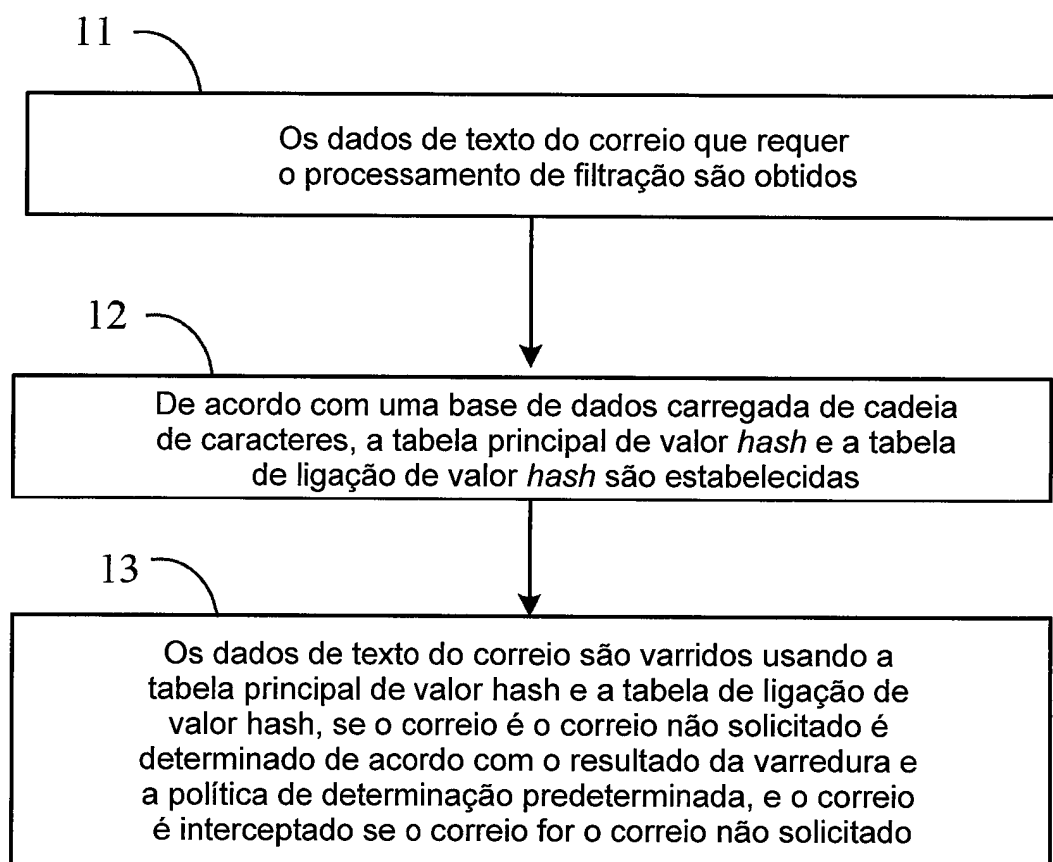


FIG. 1

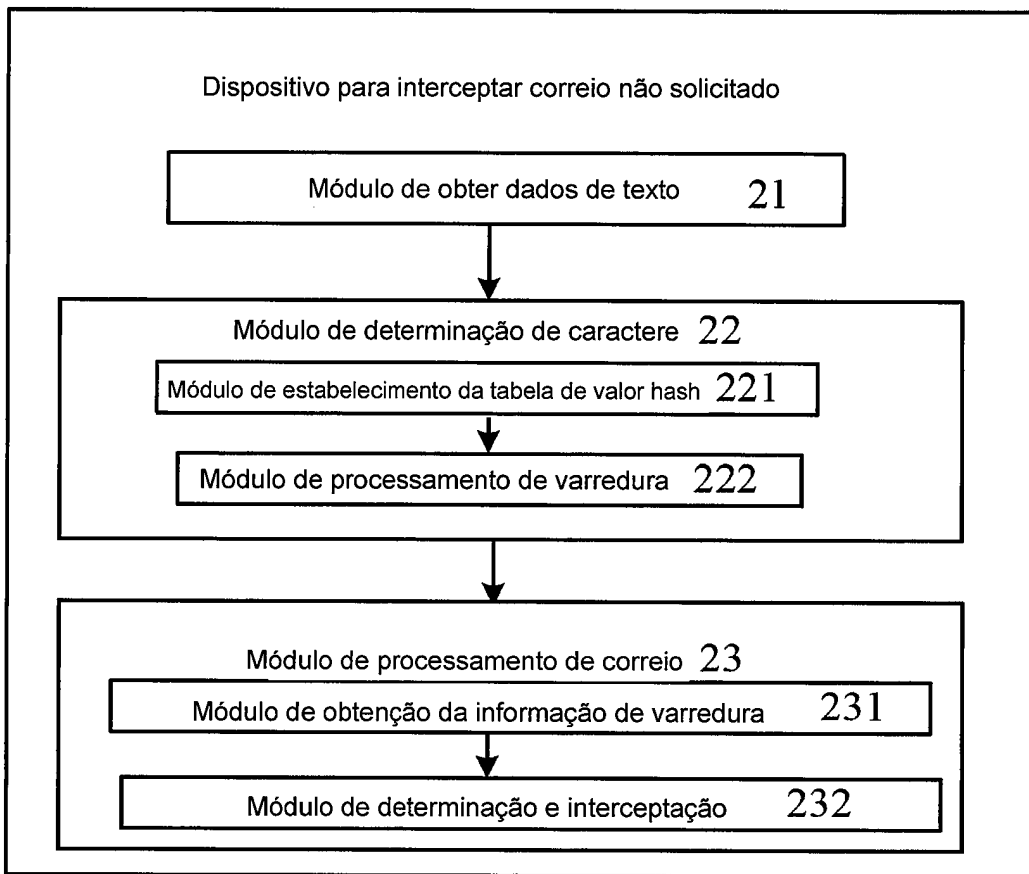


FIG. 2