(11) Número de Publicação: **PT 1315096 E**

(51) Classificação Internacional:
G06F 17/30 (2011.01)

(12) **FASCÍCULO DE PATENTE DE INVENÇÃO**

(22) Data de pedido: **2001.11.21**

(30) Prioridade(s):

(43) Data de publicação do pedido: **2003.05.28**

(45) Data e BPI da concessão: **2012.07.18**
194/2012

(73) Titular(es):
WERNER VOEGELI
QUELENSTRASSE 62 5330 ZURZACH **CH**

(72) Inventor(es):
WERNER VOEGELI **CH**

(74) Mandatário:
LUÍS MANUEL DE ALMADA DA SILVA CARVALHO
RUA VÍCTOR CORDON, 14 1249-103 LISBOA **PT**

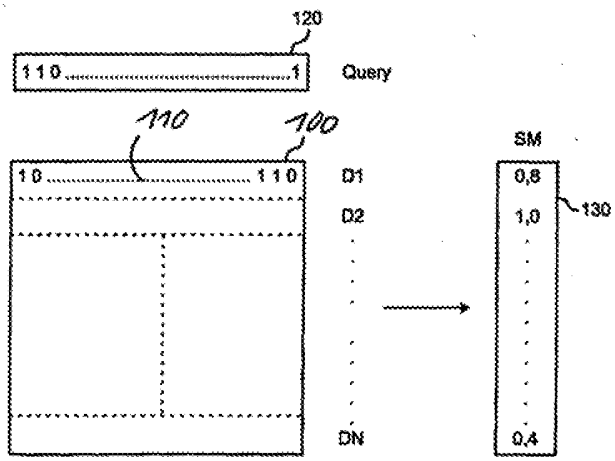
(54) Epígrafe: **MÉTODO E DISPOSITIVO PARA PESQUISAR INFORMAÇÃO RELEVANTE**

(57) Resumo:
UM MÉTODO PARA PESQUISAR A PARTIR DE UM CONJUNTO DE DOCUMENTOS ELECTRÓNICOS AQUELES DOCUMENTOS QUE ESTEJAM PRÓXIMOS EM CONTEÚDO DE UM DOCUMENTO ELECTRÓNICO UTILIZADO COMO DOCUMENTO DE ENTRADA, COMPREENDENDO O REFERIDO MÉTODO: GERAR UMA PLURALIDADE DE FRAGMENTOS DO REFERIDO DOCUMENTO DE ENTRADA QUE SÃO PARTES DO REFERIDO DOCUMENTO DE ENTRADA; PARA CADA UM DOS REFERIDOS FRAGMENTOS, UTILIZAR OS REFERIDOS FRAGMENTOS DO REFERIDO DOCUMENTO DE ENTRADA COMO UMA ENTRADA PARA UMA MEMÓRIA ASSOCIATIVA NA QUAL O REFERIDO CONJUNTO DE DOCUMENTOS ESTÁ ARMAZENADO DE FORMA A ENCONTRAR UMA MEDIDA DE SIMILARIDADE ENTRE OS FRAGMENTOS DE ENTRADA E CADA UM DOS DOCUMENTOS NA REFERIDA MEMÓRIA ASSOCIATIVA; CALCULAR UMA MEDIDA FINAL DE SIMILARIDADE PARA CADA UM DO REFERIDO CONJUNTO DE DOCUMENTOS COM BASE NAS MEDIDAS DE SIMILARIDADE OBTIDAS PARA A REFERIDA PLURALIDADE DE FRAGMENTOS; COM BASE NA REFERIDA MEDIDA FINAL DE SIMILARIDADE, DETERMINAR QUAIS DOCUMENTOS PODEM SER VISTOS COMO TENDO UM CONTEÚDO PRÓXIMO DO REFERIDO DOCUMENTO DE ENTRADA.

RESUMO

"MÉTODO E DISPOSITIVO PARA PESQUISAR INFORMAÇÃO RELEVANTE"

Um método para pesquisar a partir de um conjunto de documentos electrónicos aqueles documentos que estejam próximos em conteúdo de um documento electrónico utilizado como documento de entrada, compreendendo o referido método: gerar uma pluralidade de fragmentos do referido documento de entrada que são partes do referido documento de entrada; para cada um dos referidos fragmentos, utilizar os referidos fragmentos do referido documento de entrada como uma entrada para uma memória associativa na qual o referido conjunto de documentos está armazenado de forma a encontrar uma medida de similaridade entre os fragmentos de entrada e cada um dos documentos na referida memória associativa; calcular uma medida final de similaridade para cada um do referido conjunto de documentos com base nas medidas de similaridade obtidas para a referida pluralidade de fragmentos; com base na referida medida final de similaridade, determinar quais documentos podem ser vistos como tendo um conteúdo próximo do referido documento de entrada.


120
1 1 0 1
Query
110
100
1 0 ..... 1 1 0
D1
D2
DN
SM
0,8
1,0
0,4
130

DESCRIÇÃO

"MÉTODO E DISPOSITIVO PARA PESQUISAR INFORMAÇÃO RELEVANTE"

A presente invenção relaciona-se com um método para pesquisar, a partir de um conjunto de documentos electrónicos, aqueles documentos que estejam próximos em conteúdo de um documento electrónico utilizado como documento de entrada.

Hoje em dia, as pessoas são confrontadas com uma cada vez mais crescente quantidade de informação. A quantidade de informação produzida num único dia é tão grande que a um indivíduo, mesmo lendo 24 horas por dia durante todo o seu tempo de vida, seria impossível ler o que é produzido num único dia.

É claro que a informação que é produzida numa tão massiva quantidade tem que ser processada de alguma forma electronicamente. A forma mais convencional de processar, armazenar, arquivar e pesquisar grandes quantidades de informação são as chamadas bases de dados.

As bases de dados armazenam pedaços individuais de informação de alguma maneira ordenada que é chamada de "indexação".

Indexação significa que cada pedaço de informação

é rotulado com pelo menos uma etiqueta, sendo uma etiqueta um tipo de "meta" informação que descreve alguns aspectos do conteúdo do pedaço de informação que tenha na realidade sido armazenado.

Um exemplo muito simples é um registo de endereço que pode compreender o nome de família, nome próprio, rua, número da porta, código postal, e o nome de uma cidade. Um campo de indexação poderá então ser, por exemplo, o nome de família que descreve um aspecto da totalidade do registo de endereço.

Supondo que está armazenada uma quantidade de registos de endereços, então cada um tem uma etiqueta contendo um nome de família, e estas etiquetas poderão ser então ordenadas automaticamente e poderão ser utilizadas para aceder aos registos de endereços individuais de uma maneira rápida, como é bem conhecido na arte da tecnologia das bases de dados.

Este método é muito rápido para aceder aos registos de endereços individuais (que são chamados conjuntos de dados "datasets" na terminologia das bases de dados), no entanto, isto tem a desvantagem de que apenas podem ser pesquisados para consulta aqueles pedaços ou tipos de informação para os quais tenha sido construída uma indexação. Isto significa que o acesso a informação armazenada através de indexação não fornece necessariamente ao utilizador um quadro completo ou todos os aspectos da

informação que tenha sido armazenada, porque apenas aqueles aspectos para os quais tenha sido construída uma indexação são abertos para um utilizador para serem utilizados como um critério de pesquisa.

Uma abordagem bastante diferente é o armazenamento e o acesso à informação pela simulação de uma chamada memória associativa. De acordo com esta abordagem, cada documento de um conjunto de documentos é representado através de um assim chamado "vector de particularidades" correspondente que, de facto, é uma cadeia de bits. Cada bit da cadeia de bits assinala a presença ou a ausência de uma certa particularidade no documento que ela representa.

As particularidades podem ser, por exemplo, a presença ou ausência de certos unigramas, bigramas, trigramas, ou semelhantes.

Cada documento armazenado é então representado através de uma cadeia de bits na qual cada bit individual indica quando uma certa particularidade está presente ou ausente no seu documento correspondente. O acesso e a recuperação dos documentos assim armazenados são então levados a cabo introduzindo um documento de pesquisa. Este documento de pesquisa então é também representado através de uma cadeia de bits que é gerada de acordo com as mesmas regras das cadeias de bits que representam o documento já armazenado, isto é, para gerar um vector de particularidades para a pesquisa.

Após isto ter sido levado a cabo, é realizada uma operação lógica bit a bit entre a cadeia de bits de pesquisa e as cadeias de bits que representam os documentos já armazenados, por exemplo uma conjunção lógica "AND" bit a bit entre a pesquisa e cada um dos vectores documento armazenados. Com base nesta operação bit a bit é realizado um certo tipo de medida de similaridade entre a cadeia de bits de pesquisa de entrada e a cadeia de bits representando uma representação dos documentos que constituem a base de dados. As particularidades que estão presentes na pesquisa e no documento armazenado originam um "1" lógico como resultado da operação conjunção "AND", e quanto mais particularidades estiverem presentes na pesquisa e no documento armazenado mais "1"s lógicos resultarão da operação conjunção "AND". Como um exemplo, o número de "1"s lógicos pode ser visto como uma medida de similaridade entre uma cadeia de bits de pesquisa e a cadeia de bits (vector de particularidades) de um documento armazenado. As medidas de similaridade (individuais) obtidas desta forma para os diferentes documentos do conjunto podem ser consideradas como sendo as componentes de um vector de um vector de medida de similaridade.

Com base no vector de medida de similaridade que indica a similaridade entre o documento de pesquisa e os documentos individuais armazenados na base de dados, pode ser recuperado o documento mais similar a partir da base de dados que - em relação às particularidades codificadas na cadeia de bits - está mais próximo do documento de

pesquisa.

Com um tal acesso associativo torna-se possível introduzir qualquer pesquisa arbitrária numa base de dados e encontrar aqueles documentos que sejam mais similares ao documento de pesquisa de entrada. Uma das capacidades de uma tal pesquisa associativa é a de que aqueles documentos que são vistos como "similares" com base numa avaliação dos vectores de particularidades são na realidade similares no seu conteúdo à pesquisa de entrada. Isto significa que uma tal pesquisa associativa pode realizar um acesso verdadeiramente "associativo" que é muito similar àquilo que o cérebro humano realiza.

Por exemplo, se alguém lê um certo livro acerca de um certo tópico então um ser humano tem automaticamente uma série de associações, vêm à sua lembrança livros que se relacionam com tópicos similares ou iguais, vêm à sua lembrança eventos e experiências ligados a estes tópicos, e coisas do género. Assumindo que eles estão todos representados através de vectores de particularidades numa base de dados, então alguém poderá realizar uma pesquisa baseada, por exemplo, num livro que o utilizador esteja a ler no momento, então uma tal busca associativa electrónica será capaz de produzir associações similares tal como o cérebro humano, isto é, serão recuperados livros, documentos, e - se armazenados e codificados - experiências ou eventos similares.

Um tal método de pesquisa associativa ou "Associative Access Method (ASSA)" é conhecido por exemplo a partir de Berkovich, S., El-Qawasmeh, E., Lapir, G.: "Organization of near matching in bit attribute matrix applied to associative access methods in information retrieval", Proceedings of the 16th IASTED International Conference Applied Informatics, 23. - 25 de Fevereiro de 1998, Garmisch-Partenkirchen, Alemanha.

Correspondente, mas algo mais detalhada divulgação, é também conhecida a partir de Lapir, G. M.: "Use of associative access method for information retrieval systems" PROC. 23rd PITTSBURG CONF. ON MODELING AND SIMULATION, 1992, páginas 951 - 958, XP009099840.

Concordantemente, é conhecido um método para recuperar a partir de um conjunto de documentos electrónicos aqueles documentos que são próximos em conteúdo a um documento electrónico utilizado como um documento de entrada, que compreende:

- Proporcionar um conjunto de documentos;

- Gerar vectores de particularidades de cadeias de bits para os documentos do referido conjunto, em que cada vector de particularidades representa uma representação dos referidos documentos e indica com cada uma das suas componentes binárias do vector a presença ou ausência de uma certa particularidade dentro do respectivo

documento;

- Formar uma matriz de bits de atributos a partir dos vectores de particularidades de cadeias de bits, consistindo a matriz de bits de atributos de bits individuais e representando cada bit um certo atributo para um certo documento pela indicação da presença ou ausência de uma certa particularidade dentro do documento respectivo;

- Proporcionar um documento de entrada;

- Gerar um vector de particularidades de cadeia de bits de pesquisa representando o documento de entrada e indicando com cada uma das suas componentes binárias do vector a presença ou ausência de uma certa particularidade dentro do documento de entrada;

- Efectuar uma pesquisa associativa para o referido documento de entrada utilizando a referida matriz de bits de atributos e o vector de particularidades de cadeia de bits de pesquisa representando o documento de entrada,

- Em que a pesquisa associativa para o referido documento de entrada é efectuada pela determinação para o referido documento de entrada e cada documento do referido conjunto uma medida de similaridade individual entre o referido documento de entrada e o documento respectivo do

referido conjunto,

- Em que a medida de similaridade individual respectiva é determinada efectuando uma operação lógica bit a bit entre o vector de particularidades de cadeia de bits de pesquisa representando o documento de entrada e a cadeia de bits que representa o respectivo documento na referida matriz de bits de atributos,

Em que as referidas medidas de similaridade individuais, sendo cada uma determinada para o documento de entrada e um documento respectivo do referido conjunto e representando uma similaridade entre o referido documento de entrada e o referido documento respectivo do referido conjunto, são componentes de vectores de um vector de medida de similaridade associado ao documento de entrada e a totalidade dos documentos do referido conjunto de documentos;

- Decidindo, com base nas referidas medidas de similaridade individuais incluídas como componentes de vector no referido vector de medida de similaridade, qual documento do referido conjunto pode ser visto como tendo um conteúdo próximo ao do referido documento de entrada.

Um método adicional para recuperar, a partir de um conjunto de documentos electrónicos, aqueles documentos que sejam próximos em conteúdo a um documento electrónico utilizado como um documento de entrada é conhecido a partir

do artigo BO-REN BAI ET ALL: "Syllable-based Chinese text/spoken document retrieval using text/speech queries" INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE, AGOSTO DE 2000, WORLD SCIENTIFIC, SINGAPURA, vol. 14, N.º 5, páginas 603-616.

De acordo com esta abordagem são utilizados vectores de particularidades não binários que contêm a presença de informação, contagens de frequência e classificações acústicas de sílabas e de pares de sílabas adjacentes numa rede de sílabas. Tais vectores de particularidades gerados são armazenados numa base de dados de vectores de particularidades. Por pesquisa por texto ou por pesquisa por voz é proporcionado um documento de entrada a um subsistema de recuperação em linha. É gerado um correspondente vector de particularidades para o documento de entrada com base na mesma rede de sílabas. Com base na base de dados de vectores de particularidades e no vector de particularidades de pesquisa, um módulo de recuperação efectua uma pesquisa associativa dentro da base de dados de vectores de particularidades avaliando medidas de similaridade entre o vector de particularidades de pesquisa e os vectores de particularidades de todos os documentos da base de dados. É seleccionado um conjunto de documentos com as mais elevadas medidas de similaridade como a informação de saída da recuperação.

É conhecido um método para recuperação de dados a partir de um conjunto de documentos electrónicos com base

em “vectores de pesquisa” na forma de uma lista de identificadores de itens ou identificadores de atributos e dos seus valores correspondentes, possivelmente relacionando-se com o conteúdo de um documento electrónico utilizado como um documento de entrada, a partir da Patente dos Estados Unidos N.º 5 778 362. Uma matriz de dados bidimensional ou “mapa” que é implementado como um “quadro associativo”, pode ser organizado como uma matriz documento por termo e concordantemente pode ser baseada num conjunto de documentos que é disponibilizado. Uma entrada de célula respectiva na matriz origina a frequência com que um termo correspondente ocorre nos documentos correspondentes. Concordantemente, as linhas da matriz correspondem a vectores de particularidades que são gerados para os documentos do conjunto.

A Patente dos Estados Unidos N.º 5 778 362 divulga adicionalmente o processo de efectuar uma pesquisa associativa dentro da base de dados de vectores de particularidades com base num vector de particularidades de pesquisa. O vector de particularidades de pesquisa poderá ser baseado num documento electrónico utilizado como um documento de entrada, por exemplo uma mensagem de correio electrónico ou um documento para ser comparado com outros documentos para identificar outras similaridades.

A Patente EP 0 750 266 A1 divulga a utilização de vectores de particularidades (vectores de particularidades) de documentos de documentos de entrada. Utilizando

identificadores conceptuais que são independentes da linguagem utilizada, são gerados vectores de particularidades conceptuais que se relacionam com identificadores conceptuais em vez de identificadores baseados na língua ou baseados em fonética.

A utilização de vectores de particularidades para classificar e hierarquizar a similaridade entre ficheiros de áudio individuais está divulgada na Patente dos Estados Unidos N.º 5 918 223. Os vectores de particularidades relacionam-se com particularidades que descrevem as particularidades do som do respectivo ficheiro de áudio tais como altura, brilho e ritmo.

A utilização de vectores de particularidades representando particularidades de documentos respectivos é também conhecida a partir da Patente EP 1 049 030 A1. As componentes do vector de um vector que represente um certo documento correspondem à frequência da ocorrência de termos no referido documento. Um esquema de classificação relaciona-se com a separação entre vectores de particularidades respectivos num espaço vectorial formado pelas n dimensões dos vectores de particularidades, com subespaços neste espaço vectorial correspondentes a uma certa classe de documentos.

Em referência novamente às acima mencionadas duas primeiras citações da arte antecedente, pode existir um problema de que a resolução da pesquisa associativa

convencionalmente conhecida está limitada, dado que particularidades para as quais o vector de particularidades de pesquisa e os vectores de particularidades dos documentos se relacionam, não fornecem informação útil mas apenas uma quantidade de "ruído". Isto pode ser em particular o caso quando a dimensão dos vectores de particularidades é relativamente elevada e os documentos do conjunto bem como o documento de entrada contenham muitos aspectos diferentes e uma grande quantidade de texto. Por exemplo, quando estes documentos, em particular o documento de entrada, são/é muito extenso, muitas das particularidades sobre as quais a codificação binária das particularidades está baseada podem frequentemente ser encontradas algures no texto, de forma que a definição do bit respectivo no respectivo vector de particularidades não será uma boa base para a pesquisa associativa para encontrar documentos similares.

Para aumentar a resolução da pesquisa associativa ou para evitar uma deterioração da resolução da pesquisa associativa a invenção proporciona o método definido na Reivindicação 1.

De acordo com a invenção o documento de pesquisa de entrada não apenas é dividido em fragmentos, mas os fragmentos individuais são então deslocados de um dado número de unidades com base no qual o documento de pesquisa é composto. Esta utilização de uma chamada janela deslizante cria então fragmentos adicionais que podem ser

utilizados como entrada para a pesquisa associativa. Estas entradas adicionais originam de novo medidas de similaridade adicionais que são também tomadas em conta quando elas são combinadas numa medida final de similaridade.

A invenção está baseada na compreensão de que tais fragmentos de documento obtidos com base no documento de entrada são geralmente mais centrados a respeito do assunto ou do aspecto com que se relacionam, e concordantemente mais centrados a respeito de particularidades representadas pelas respectivas componentes de vector dos vectores de particularidades de pesquisa gerados para os fragmentos.

Concordantemente, as medidas de similaridade obtidas para os fragmentos do documento de entrada em relação à totalidade dos documentos do conjunto permitirão uma melhor decisão sobre quais documentos do conjunto podem ser vistos como tendo um conteúdo próximo ao do documento de entrada, de forma que as medidas de similaridade finais individuais incluídas como componentes de vector no vector de medida de similaridade final podem ter uma melhor resolução.

Adicionalmente, podem ser obtidas vantagens e melhorias pelos ensinamentos dados nas sub-reivindicações.

De acordo com uma particular forma de realização,

a pesquisa e os documentos do conjunto de documentos armazenados representam textos lidos por uma voz humana.

De acordo com uma particular forma de realização adicional, a pesquisa associativa pode ser utilizada para classificar o documento de pesquisa de entrada como pertencendo a uma de uma pluralidade de classes pré-definidas. Para este propósito, os documentos do conjunto são respectivamente classificados como pertencendo a uma de uma pluralidade de classes, depois é realizada a pesquisa associativa à medida que as concordâncias com as classes individuais são contadas. A classe para a qual tenham sido encontradas mais concordâncias, é então a classe com a qual o documento de pesquisa de entrada é associado.

De acordo com uma particular forma de realização adicional é utilizado um método com auto-optimização para optimizar os parâmetros da pesquisa associativa. Para este propósito, é utilizada uma base de dados ideal com duas ou mais classes que não contêm classificações incorrectas. O método de classificação de acordo com uma forma de realização da invenção é então utilizado para tentar reproduzir esta classificação ideal classificando automaticamente todos os documentos contidos na base de dados. Este processo é levado a cabo repetidamente enquanto se variam sistematicamente os parâmetros do processo. Aqueles métodos que não fornecem bons resultados e para os quais os parâmetros aparentemente não estejam optimizados

não são mais utilizados. Finalmente, será alcançado um conjunto optimizado de parâmetros que reproduz a base de dados ideal de uma forma óptima através da classificação automática.

Descrição Detalhada.

Uma pesquisa associativa pela simulação de uma memória associativa é, em princípio, agora explicada em ligação com a Figura 1. Para a pesquisa associativa é armazenado um conjunto de documentos D1 a DN. Cada um destes documentos é representado através de um correspondente vector de particularidades 110 que é uma sequência de um pré-determinado número de bits. Cada um destes bits representa ou indica a presença ou ausência de uma certa particularidade no documento que ele representa. Uma tal particularidade pode, por exemplo, ser um trigrama e o vector de particularidades pode então ter um comprimento em concordância com o possível número de trigramas. Evidentemente, a representação de outras particularidades é também possível.

Como já mencionado, cada documento é representado através de um correspondente vector de particularidades 110. Estes vectores de particularidades formam juntos uma matriz de vectores de particularidades 100 que também pode ser chamada de matriz de bits de atributos devido a ela consistir de bits individuais e cada bit representar um certo atributo para um certo documento.

Numa maneira análoga, não apenas os documentos estão representados, mas também o documento de pesquisa pode ser representado através de uma cadeia de bits 120 que é formada de acordo com as mesmas regras que são válidas para a formação da matriz de bits de atributos. De acordo com uma abordagem da arte antecedente, uma tal cadeia de bits de pesquisa é gerada para representar o documento de pesquisa. A cadeia de bits de pesquisa 120 é então combinada com a matriz de bits de atributos 100 pela realização de operações lógicas bit a bit de uma maneira tal que para cada um dos documentos D1 a DN é obtida uma medida de similaridade que é então armazenada num vector de medida de similaridade 130. A forma mais simples para calcular a medida de similaridade é a realização de uma operação lógica de conjunção "AND" entre a cadeia de bits de pesquisa e cada uma das linhas na matriz 100. Isto dá então para cada linha na matriz 100 uma linha resultado correspondente que tem um certo número de "1" lógicos. Quanto maior for o número de "1" lógicos, maior é a similaridade entre a cadeia de bits de pesquisa e um certo vector de particularidades. Com base nisto pode ser calculada a medida de similaridade. Podem também ser imaginadas outras maneiras, mais sofisticadas, para calcular a medida de similaridade, por exemplo para além de apenas realizar a operação sobre os bits é também possível ter em conta os conteúdos reais dos documentos D1 a DN, similaridades fonéticas entre a cadeia de bits de pesquisa e os respectivos documentos. Estas similaridades fonéticas podem também ser tomadas em conta quando se calcula a

medida de similaridade final.

De acordo com a invenção, são geradas não apenas uma cadeia de bits de pesquisa representando a totalidade do documento de pesquisa mas várias cadeias de bits de pesquisa, cada uma representando um fragmento respectivo do documento de pesquisa, como explicado no que se segue.

A Figura 2 explica como pode ser processada uma pesquisa de entrada de forma a se obter um acesso associativo ou pesquisa associativa mais eficientes. A pesquisa de entrada 200 é primeiro separada em fragmentos S1, S2, e S3 com uma dada dimensão. A dimensão dos fragmentos é um parâmetro que pode tanto ser pré-definido como pode ser seleccionado pelo utilizador ou pode ser definida automaticamente utilizando um método optimizado que é utilizado para encontrar um conjunto de parâmetros optimizado. A separação está baseada em unidades de elementos baseados naquilo do qual o documento de pesquisa é compreendido, tal como por exemplo baseado em palavras. O documento de pesquisa de entrada pode por exemplo compreender 15 palavras, e a dimensão dada do fragmento pode ser cinco, então os segmentos S1, S2, e S3 são cada um com cinco palavras de comprimento. Se a pesquisa compreender apenas 14 palavras, então o último fragmento S3 conterá apenas quatro palavras.

Os segmentos individuais S1 a S3 são então convertidos em correspondentes vectores de particularidades

F1 a F3, como indicado na Figura 2. Então, para cada um deles, é realizada uma pesquisa associativa tal como aquela explicada em ligação com a Figura 1.

Dado que existem três vectores de pesquisa de entrada, concordantemente também haverá três vectores de medida de similaridade ou vectores de similaridade que são o resultado da pesquisa. As componentes de vector dos vectores de similaridade representam a similaridade entre a pesquisa de entrada, aqui o respectivo fragmento para o qual o respectivo vector de pesquisa de entrada foi codificado, e um documento individual respectivo do conjunto armazenado. Cada um dos vectores de medida de similaridade SM1 - SM3 representa um aspecto de similaridade entre o documento de pesquisa de entrada, a partir do qual os fragmentos foram obtidos, e o conjunto de documentos estando armazenados por uma respectiva cadeia de bits para a pesquisa associativa.

De acordo com a invenção, são para serem obtidos vectores de pesquisa de entrada adicionais gerando fragmentos adicionais utilizando um chamado “deslocamento”, que é explicado abaixo.

Com base nos vectores de medida de similaridade individuais SM1 a SM3 é então calculado um vector de medida de similaridade final FSM como se mostra na Figura 3. A Figura 3 ilustra esquematicamente o vector adição dos vectores de medida de similaridade individuais SM1 a SM3.

Deverá ser aqui salientado que a operação para se obter o vector de medida de similaridade final FSM pode também compreender uma ponderação em concordância com uma certa função de ponderação. Por exemplo, aquelas componentes de vector dos vectores de medida de similaridade SM1 a SM3 que indicam uma maior similaridade podem ser ponderadas mais fortemente do que aquelas que tenham uma similaridade inferior (por exemplo abaixo de 50%). Isto atribui um peso maior àqueles documentos do conjunto armazenado para os quais tenha sido, no caso, encontrada uma similaridade e de forma a diminuir o peso daqueles documentos para os quais a similaridade seja, no caso, apenas um tipo de "ruído".

Uma forma de realização da invenção é agora explicada em conexão com a Figura 4. Assuma-se que um documento de pesquisa de entrada tem 16 palavras de comprimento. Assuma-se ainda que a dimensão dos fragmentos é de quatro palavras, então a pesquisa de entrada 400 é separada em quatro blocos 410 a 440 como se mostra na Figura 4, tendo cada bloco quatro palavras.

Adicionalmente a esses quatro fragmentos de entrada que são gerados pela separação da pesquisa de entrada, são gerados fragmentos adicionais por um processo de deslocamento (janela deslizante) como é também ilustrado na Figura 4. Um tal deslocamento significa que as extremidades do bloco gerado através da separação da pesquisa de entrada são deslocadas de uma certa quantidade

de elementos, aqui de duas palavras. Isto resulta em quatro blocos de fragmentos 450 a 480 como ilustrado na Figura 4, onde as extremidades dos blocos individuais estão deslocadas em comparação com os quatro blocos originais.

Isto está ilustrado adicionalmente na Figura 5, onde os blocos individuais e o número de palavras neles contidas estão apresentados esquematicamente. A separação em fragmentos gera quatro blocos 500, contendo cada um quatro palavras, e o deslocamento gera quatro blocos adicionais, em que o último bloco nestes quatro blocos 510 apenas contém duas palavras, nomeadamente a palavra 15 e 16 da pesquisa de entrada original. Como consequência, através da separação e do deslocamento são gerados no total oito fragmentos da pesquisa de entrada que podem então ser utilizados como entrada para uma pesquisa associativa como explicado anteriormente. Consequentemente, não há apenas três vectores de medida de similaridade como na Figura 2, mas em vez disso haverá até oito vectores de medida de similaridade. Com base nestes oito vectores de medida de similaridade será calculado um vector de medida de similaridade final de uma maneira análoga à explicada em conexão com a Figura 3.

Com base nas medidas de similaridade finais assim obtidas como componentes de vector do vector de medida de similaridade final é realizada a recuperação. Para cada componente no vector de medida de similaridade final FSM existe um correspondente documento armazenado no conjunto,

e dependendo do conjunto de critérios será recuperado ou apenas o documento mais relevante, ou serão recuperados aqueles documentos cuja medida de similaridade caia além de um certo valor limite, ou serão recuperados os 10 documentos mais relevantes ou semelhante. O tipo de critérios de recuperação pode ser alterado dependendo do desejo do utilizador, será recuperado como informação de saída, nenhum, um ou mais documentos.

Devido à recuperação ter sido executada avaliando a similaridade entre o documento de pesquisa de entrada e os documentos armazenados para acesso associativo, a informação de saída pode ser tal que seja similar à pesquisa de entrada em relação aos vectores de particularidades.

Dado que os vectores de particularidades, ou por outras palavras, as particularidades propriamente ditas representam o conteúdo real da pesquisa de entrada bem como dos documentos do conjunto, não existe apenas uma similaridade nestas particularidades mas, na realidade, também uma similaridade em conteúdo entre aqueles documentos recuperados para os quais existe uma elevada medida de similaridade. Isto significa que com o método explicado aqui anteriormente, um utilizador pode gerar ou aplicar uma pesquisa de entrada arbitrária que defina o campo de interesse para o qual o utilizador deseja levar a cabo algumas “associações”, e a pesquisa associativa na realidade levará a cabo tais associações disponibilizando

aqueles documentos que sejam similares em conteúdo ao documento de pesquisa de entrada.

A separação do documento de pesquisa de entrada em certo sentido “aumenta a resolução” do acesso à memória associativa. Mais do que isso, dado que o acesso à memória associativa realiza “associações verdadeiras” em certo sentido, é vantajoso se a pesquisa de entrada estiver centrada num certo assunto ou num certo aspecto de um tópico em vez de conter muitos aspectos diferentes e uma grande quantidade de texto. Isto é alcançado pela separação da pesquisa de entrada em fragmentos individuais. Aqueles fragmentos são mais centrados, e eles originam uma imagem bastante boa e centrada da similaridade entre os documentos representados por vectores de particularidades de cadeia de bits na matriz de atributos de bits para o acesso associativo e os fragmentos individuais em vez de disponibilizar uma medida de similaridade não centrada que contém uma grande quantidade de “ruído”, no caso da pesquisa de entrada ser muito comprida.

De forma a evitar que o ruído seja então de novo introduzido novamente quando se calcula o vector de medida de similaridade final o vector de medida de similaridade final é calculado utilizando alguma função de ponderação que atribua mais peso àqueles documentos para os quais tenha sido medido um valor de similaridade mais elevado do que para aqueles documentos para os quais tenha sido medido um valor de similaridade mais baixo. Um exemplo para uma

tal função de ponderação está ilustrado esquematicamente na Figura 6. Como pode ser observado a partir da Figura 6, àqueles documentos para os quais a medida de similaridade está acima de 50% será atribuído um peso maior, e àqueles para os quais a medida de similaridade esteja abaixo de 50% será atribuído um peso menor. Perto de 100% e perto de 0% a função de ponderação é bastante inclinada de maneira a dar um peso elevado àqueles fragmentos da pesquisa de entrada para os quais a correspondente medida de similaridade para os documentos respectivos tenha sido calculada como sendo relativamente alta. Quando se calcula a medida de similaridade final como se ilustra na Figura 3, a função de ponderação é tomada em conta de uma maneira que para cada uma das componentes dos vectores de medida de similaridade individuais SM1 a SM3 o valor de ponderação é tomado a partir da função de ponderação apresentada na Figura 6, e a componente do vector é multiplicada pelo valor de ponderação correspondente. Desse modo pode ser atribuído um peso mais alto a valores de medida de similaridade mais elevados quando se calcula o vector de medida de similaridade final FSM.

O efeito que é alcançado pela separação da pesquisa em fragmentos e pelo deslocamento dos fragmentos, como explicado anteriormente, pode também ser alcançado de uma maneira similar por um método representando uma forma de realização da invenção como explicado aqui abaixo. A Figura 7 apresenta um vector de particularidades de pesquisa de entrada 700 que tem N elementos individuais,

aqui palavras. Com base numa dimensão dos fragmentos n que é ou pré-definida ou seleccionada ou introduzida por um utilizador, é gerado um primeiro fragmento 710. Então, de acordo com um parâmetro adicional que pode ser chamado um parâmetro de deslocamento são gerados s fragmentos adicionais da maneira seguinte. Um primeiro fragmento é gerado a partir da palavra número s e estendendo-se até ao número de palavra n+s. Este fragmento está apresentado como fragmento 720 na Figura 7. É então gerado um fragmento adicional a partir da palavra número 2s e que se estende até à palavra número n+2s, mantendo dessa forma uma dimensão do fragmento de n.

Desta maneira são gerados múltiplos fragmentos deslocando a palavra de partida para cada fragmento de s palavras para o lado direito do documento de pesquisa de entrada 700. Isto resulta numa pluralidade de fragmentos da pesquisa de entrada, e para cada um destes fragmentos é então gerado um correspondente vector de particularidades que é utilizado como pesquisa de entrada para a pesquisa associativa.

O que as formas de realização descritas até aqui têm em comum é que todas elas utilizam múltiplos fragmentos gerados com base num documento de pesquisa de entrada, e com base nestes múltiplos fragmentos são gerados múltiplos vectores de particularidades de pesquisa que podem ser utilizados para efectuar a pesquisa associativa. Os parâmetros individuais tais como a dimensão dos fragmentos

individuais ou o valor ou número pelos quais os fragmentos individuais se sobrepõem, tal como o valor s na Figura 7, podem ou ser pré-definidos de acordo com um conjunto adequado de parâmetros, eles podem ser seleccionados ou escolhidos pelo utilizador, ou eles podem ser optimizados através de determinado procedimento que deriva os parâmetros e encontra um conjunto de valores que origina um resultado óptimo, como também já explicado anteriormente.

De acordo com uma forma de realização adicional não apenas o documento de pesquisa de entrada é utilizado para gerar fragmentos, mas também os documentos armazenados na memória associativa são utilizados para gerar fragmentos de uma maneira similar à fragmentação do documento de pesquisa de entrada. Se uma tal medida é tomada, então isto resulta na realidade em múltiplas matrizes de atributos de bits, que, com certeza, aumentam a potência de cálculo necessária para levar a cabo um acesso associativo. Se tais múltiplas matrizes de atributos de bits tivessem sido geradas pela geração de múltiplos fragmentos dos documentos armazenados na memória associativa e se calculassem os correspondentes vectores de particularidades, então os fragmentos da pesquisa de entrada são aplicados como pesquisas para a totalidade das matrizes de atributos de bits, dessa forma aumentando o número de medidas de similaridade obtidas. Se uma tal abordagem será realizável em termos de potência de cálculo poderá depender do computador utilizado bem como do número de documentos que na realidade têm que estar armazenados e representados.

Com respeito ao deslocamento dos fragmentos individuais de forma a ter fragmentos que se sobreponham como explicado em ligação com a Figura 7, um tal deslocamento pode ser levado a cabo com base em qualquer das unidades com base nas quais o documento de pesquisa seja na realidade composto. Uma tal unidade pode por exemplo ser uma palavra, então isto poderá significar que no caso da Figura 7 como já explicado o primeiro fragmento contém n palavras, o segundo fragmento começa com a palavra de ordem s e estende-se até à palavra de ordem (n+s), e por aí fora. Uma outra unidade poderá também ser uma sílaba em vez de uma palavra, ou até um carácter. Qualquer unidade pode ser utilizada que seja uma unidade baseada na qual o documento de pesquisa seja composto. No entanto, dado que o processo de acesso à memória associativa simulada é acerca de obter associações com de alguma forma façam sentido em termos do seu conteúdo informativo, parece ser razoável que os elementos utilizados para gerar os fragmentos e para gerar fragmentos que se sobreponham deverão ser elementos que eles próprios contenham algum tipo de significado ou pelo menos alguma informação neles próprios, tais como palavras ou pelo menos sílabas. Uma outra de tais unidades poderá ser um fonema, isto será explicado em mais detalhe em ligação com uma forma de realização adicional.

Agora será explicada, em ligação com a Figura 8, uma forma de realização da presente invenção que pode ser utilizada para classificar um documento de pesquisa de

entrada de acordo com qual de uma pluralidade de classes de classificação ele pertence. Na Figura 8 está apresentada a matriz de atributos de bits 800 representando um número N de documentos D_1 a D_n pelos respectivos vectores de particularidades de cadeia de bits. Cada um destes documentos está classificado com pertencendo a uma de três classes de classificação. Por exemplo, os documentos D_1 a D_n podem ser artigos, e eles podem ser classificados como sendo quer artigos científicos (classe 1), artigos de imprensa (classe 2), ou outros artigos (classe 3). O vector 810 apresentado na Figura 8 contém o rótulo que atribui uma classe de classificação a cada um dos documentos D_1 a D_n.

Permita-se-nos agora assumir que o número total de documentos N é muito maior do que 3, de forma a que cada uma das classes de classificação contém um número comparativamente grande de documentos. Permita-se-nos agora ainda assumir que um documento de pesquisa de entrada 820 seja utilizado como pesquisa de entrada para a pesquisa associativa. Um utilizador pode agora estar interessado não apenas em recuperar aqueles documentos de entre D_1 a D_n que sejam mais similares à pesquisa 820, mas ele pode também estar interessado em classificar se o documento de pesquisa 820 pertence a qual das classe 1, classe 2, ou classe 3.

Permita-se-nos agora assumir que a pesquisa é levada a cabo para uma pluralidade de fragmentos, e permita-se-nos assumir ainda que o critério de recuperação é tal que sejam recuperados todos os documentos para os

quais a medida de similaridade esteja acima de um certo limite. Se o limite de similaridade for definido de tal modo que seja obtida uma pluralidade de resultados de recuperação, então haverá uma pluralidade de concordâncias (documentos recuperados) para cada uma das diferentes classes 1, 2, ou 3. No entanto, com base na assunção que os documentos que pertençam a uma certa classe são de alguma forma auto-similares, haverá um número significativamente maior de documentos recuperados a partir da classe da qual o documento de pesquisa 820 na realidade também pertença.

A Figura 9 ilustra as estatísticas de recuperação que podem ser obtidas quando se realiza uma pesquisa baseada num documento de pesquisa de entrada sobre um conjunto de documentos armazenados, isto é, documentos representados por um respectivo vector de particularidades de cadeia de bits na matriz de atributos de bits, que pertence a uma das três classes. No exemplo apresentado na Figura 9 existem 85 concordâncias que pertencem à classe 2, isto significa que o documento de pesquisa de entrada muito presumivelmente também pertence a esta classe. Uma "concordância" significa aqui que um resultado de pesquisa satisfaz um certo critério tal como uma medida de confidencialidade que está acima de um certo limite de forma a que ela pode ser vista como uma "concordância". O critério particular e o valor do limite podem ser seleccionados dependendo das circunstâncias.

Com base numa tal estatística de contagem de

concordâncias pode ser decidido sobre qual classe de um conjunto de classes um documento de entrada pertence, e pode ser levada a cabo uma classificação automática. Esta classificação automática será tanto melhor quanto mais precisa for já a classificação original. Se o vector de classificação 810 da Figura 8 já contém algumas classificações incorrectas, então também as estatísticas de recuperação da Figura 9 conterão erros correspondentes. No entanto, se a classificação original for bastante boa, então a probabilidade será também relativamente elevada de que as estatísticas de recuperação sejam uma boa base para classificar um documento de pesquisa desconhecido e não classificado.

Nesta conexão, deverá também ser referido que o conjunto de documentos para os quais a classificação original seja muito boa é o mais adequado como conjunto de partida para se encontrar um conjunto optimizado de parâmetros tais como a dimensão da fragmentação e o valor do deslocamento ou sobreposição explicado em ligação com a Figura 7. Permita-se-nos assumir que é utilizada uma muito boa classificação e um correspondente conjunto de documentos já classificados, então para cada um dos documentos neste conjunto é realizada uma classificação automática como explicado anteriormente. Dado que a classificação original já é boa ou num caso ideal é já perfeita, as estatísticas de recuperação da Figura 9 deverá dar sempre o valor correcto no sentido de que a classificação resultante deverá também ser correcta. Se

este não for o caso, então possivelmente alguma coisa está errada com os parâmetros, e os parâmetros deverão ser alterados. Com os parâmetros alterados então pode outra vez ser realizado um novo processamento com base no conjunto de dados original, e pode ser verificado se a classificação com os novos parâmetros dá agora um melhor resultado. Com um tal método de uma maneira optimizada pode ser encontrado um conjunto de parâmetros ideal ou óptimo.

Até agora não foi prestada atenção ao tipo de documentos de entrada utilizados como documento de pesquisa para pesquisar a memória associativa simulada. Em princípio, pode ser utilizado qualquer tipo de documento de entrada, e qualquer tipo de documento pode ser representado por um vector de particularidades para a pesquisa associativa. No entanto, daqui em diante será explicada uma forma de realização particular onde o documento de pesquisa de entrada bem como os documentos representados para o acesso associativo são documentos de voz.

Permita-se-nos assumir que o conjunto de documentos a serem representados para a pesquisa associativa são ficheiros que representam alguns textos falados por uma voz humana. Estes textos podem por exemplo ser discursos radiofónicos, discursos televisivos, entrevistas, conversações telefónicas, ou semelhantes. Estes documentos de voz podem ser convertidos em documentos de texto de acordo com um método convencional de voz-para-texto resultando então num correspondente ficheiro

de texto. Possivelmente, antes da conversão voz-para-texto poderá ter sido realizada uma gravação da voz que grave a voz em algum ficheiro de som, tal como um ficheiro .wav. Isto está ilustrado esquematicamente na Figura 10.

Podem então haver por exemplo muitos de tais ficheiros de texto que são resultantes de alguma voz humana, e esses ficheiros de texto poderão então ser armazenados como documentos numa memória associativa simulada através dos seus correspondentes vectores de particularidades. Para este propósito, cada um dos ficheiros de texto é convertido no seu correspondente vector de particularidades, e o assim resultante conjunto de vectores de particularidades é organizado numa matriz para se obter a matriz de atributos de bits como já explicado em ligação com a Figura 1.

É então possível utilizar também algum texto falado por uma voz humana como uma pesquisa de entrada. Para este propósito, similar ao processo apresentado na Figura 10 é gerado um ficheiro de texto representando a pesquisa. Da mesma forma este ficheiro de texto é representado num correspondente vector de particularidades, e então pode ser obtido um acesso à memória associativa como já explicado anteriormente.

Utilizando um tal método, torna-se possível recuperar a partir de um arquivo de som aqueles itens que sejam similares em conteúdo a qualquer pesquisa de som

dada. Permita-se-nos por exemplo assumir que os documentos representados para o acesso associativo por um respectivo vector de particularidades são texto de notícias, então será possível dizer um texto de busca para um microfone, gravá-lo de acordo com a Figura 10, transferi-lo para um ficheiro de texto, e recuperar a partir da memória associativa simulada aqueles ficheiros de texto e então os correspondentes ficheiros de voz que sejam mais similares em conteúdo com a pesquisa ditada.

Uma forma de realização adicional da presente invenção será agora explicada em detalhe em ligação com a Figura 11. Numa base de dados 900 estão armazenadas amostras de ficheiros de áudio para permitir um acesso associativo. Para este propósito, as amostras foram convertidas em ficheiros de texto, os ficheiros de texto foram convertidos em vectores de particularidades, e os vectores de particularidades foram organizados numa matriz de atributos de bits como já explicado.

Um ficheiro de áudio 910, que pode, por exemplo, tomar a forma de um ficheiro .wav, é então aplicado ao sistema e pré-processado acusticamente no passo 920. O problema com ficheiros acústicos é que algumas vezes a gama dinâmica é demasiado grande, o nível de ruído é demasiado elevado, etc. Para melhorar a qualidade no passo de processamento 920, a gama dinâmica é normalizada, o ruído é filtrado, etc. Isto assegura que no seguinte passo 930 do reconhecimento de voz o motor de reconhecimento de voz

recebe um ficheiro de áudio que está mais adequado para obter os melhores resultados no processo de conversão de voz-para-texto. O passo 930 retorna como saída de informação um texto no qual o ficheiro de áudio tinha sido convertido.

No passo 940 um sequenciador/analisador gramatical é então utilizado para gerar os fragmentos da amostra de texto de entrada. Isto pode ser levado a cabo de acordo com qualquer dos métodos para gerar fragmentos explicado anteriormente. Um resultado do sequenciador/analisador gramatical é uma pluralidade de fragmentos de textos do texto gerado pela unidade de reconhecimento de voz 930.

Com base nesta pluralidade de fragmentos de texto é então levado a cabo o acesso associativo directo no passo 950 para cada um dos fragmentos. Isto resulta então numa respectiva medida de similaridade para as amostras armazenadas por um respectivo vector de particularidades na memória da base de dados 900, e dependendo do critério de saída de informação definido é colocada na saída uma ou mais amostras tendo a mais elevada medida de similaridade, ou tendo uma medida de similaridade acima de um certo limite. Aquelas amostras colocadas na saída de informação podem então ser consideradas como sendo as mais similares à amostra que tenha sido colocada na entrada de informação como um ficheiro de áudio no passo 910.

De acordo com uma forma de realização adicional, o método explicado em ligação com a Figura 11 pode ser ainda mais refinado tendo dois conjuntos de amostras, um conjunto de amostras 900 como se apresenta na Figura 11, e um conjunto adicional de amostras 905 que não está apresentado na Figura 11. Permita-se-nos agora assumir que o primeiro conjunto de amostras 900 contém amostras relevantes no sentido de que elas são amostras que podem ser classificadas como sendo interessantes do ponto de vista do utilizador, e o outro conjunto de amostras 905 contém amostras de voz ou amostras de ficheiros de voz que não sejam interessantes do ponto de vista do utilizador. Num tal caso, o acesso a memória associativa pode ser combinado com uma estatística de recuperação como explicado em ligação com a Figura 9. Para um dado ficheiro de áudio de entrada 910 haverá um número de amostras recuperadas a partir do conjunto 900 e um outro número de concordâncias no conjunto de amostras 905. Se o conjunto relevante de amostras 900 fornecer o maior número de concordâncias, então o ficheiro de áudio aplicado na entrada de informação pode ser assumido como tendo também relevância para o utilizador. No entanto, se o número de concordâncias no conjunto não relevante de amostras 905 é maior, então pode ser assumido que o ficheiro de áudio no seu conteúdo não tem interesse particular para o utilizador e não necessita de ser colocado na saída de informação conjuntamente com as concordâncias encontradas. Com um tal método, a classificação e recuperação podem ser combinadas de maneira a verificar um ficheiro de áudio de entrada desconhecido se

ele é relevante de todo, e se ele é relevante de todo, para recuperar aquelas amostras já armazenadas que sejam similares em conteúdo ao ficheiro de áudio aplicado na entrada de informação.

De acordo com uma forma de realização adicional no passo no passo 930 de reconhecimento de voz a conversão não é realizada para um ficheiro de texto, mas em vez disso para um ficheiro cujos elementos são fonemas codificados de forma apropriada, tal como através de sequências de caracteres ASCII. Deverá ser aqui referido que o reconhecimento de voz contém maioritariamente dois passos, primeiro uma conversão dos ficheiros áudio em fonemas (um reconhecimento dos fonemas das palavras individuais faladas), e então uma conversão a partir dos fonemas reconhecidos para as palavras finais no texto. Ocorre relativamente muitas vezes no processamento do reconhecimento de voz que os fonemas são correctamente reconhecidos, no entanto, o reconhecimento final das palavras individuais não é realizado correctamente. Se o passo 930 de reconhecimento de voz está limitado para realizar apenas a conversão em fonemas, e se a base de dados de amostras 900 também armazena a amostra na forma de fonemas em vez de na forma de texto, então esta fonte de erros pode ser evitada e pode ser esperado um processamento mais preciso.

Como será evidente a partir da explicação precedente, o sistema explicado em ligação com a Figura 11

pode ser utilizado para dois propósitos. Primeiramente, no caso do utilizador ter uma certa imaginação sobre o conteúdo sobre o qual ele queira que a associação tenha, ou por outras palavras, se ele já sabe qual o tipo de conteúdo que deverão ter os documentos que deverão ser recuperados, um utilizador pode apenas compor um ficheiro de áudio de entrada recitando um certo texto que ele considere como interessante ou relevante. O processo apresentado na Figura 11 recuperará então como saída de informação aqueles documentos de entre as amostras armazenadas que são mais similares no seu conteúdo ao ficheiro de áudio aplicado na entrada de informação gerado pelo utilizador.

Uma outra aplicação poderá ser a de que o conteúdo e a relevância do ficheiro de áudio 910 propriamente dito não sejam ainda conhecidas. Realizando um acesso associativo directo como explicado em ligação com as duas amostras 900 e 905, pode ser então primeiramente verificado se o ficheiro de áudio é de todo relevante a partir do ponto de vista do utilizador verificando se a sua classificação o classificará como sendo relevante ou não relevante, por outras palavras como pertencendo às amostras 900 ou às amostras 905. Se o ficheiro aplicado na entrada de informação for avaliado como sendo relevante, então pode ser realizada uma recuperação por acesso associativo e documentos similares podem ser apresentados na saída de informação para o utilizador a partir do conjunto de amostras 900.

O método descrito em ligação com a forma de realização precedente é uma ferramenta muito poderosa que pode ser utilizada para reduzir “associações” como as do cérebro humano. Com base num documento de pesquisa aplicado na entrada de informação são produzidos documentos de saída que são associações relacionadas com o documento de entrada de uma maneira tal que eles têm um conteúdo similar ao do documento de entrada. “Conteúdo similar”, aqui, é bastante difícil de descrever em termos mais exactos, uma outra tentativa poderá ser dizer que os documentos encontrados têm aspectos sob os quais eles possam ser vistos como similares ao documento de pesquisa aplicado na entrada de informação. Estes aspectos dependam em alguma extensão da forma como o vector de particularidades é produzido com base no documento de pesquisa e dos documentos armazenados para o acesso associativo. Se os vectores de particularidades são produzidos com base em bigramas ou trigramas, então os documentos recuperados representam similaridades em termos de sílabas consecutivas. Isto também reflecte necessariamente alguma similaridade em conteúdo, porque palavras similares que ocorram no documento de pesquisa e nos documentos armazenados resultarão em vectores de particularidades similares ou idênticos.

No entanto, dependendo de como a medida de similaridade seja calculada na realidade, também outros aspectos de similaridade podem ser tomados em conta, tais como similaridade fonética que poderá ser bastante

significativa se os fonemas forem utilizados para construir os vectores de particularidades em vez de sequências de caracteres ASCII.

A presente invenção nas suas formas de realização é capaz, portanto, de realizar verdadeiras "associações", e isto é uma ferramenta muito poderosa para qualquer ser humano que seja confrontado com o processamento do dilúvio de informação que ele vai encontrando a cada dia. Para muitas ou para a totalidade das peças de informação que encontramos hoje em dia temos que levar a cabo a tarefa de realizar associações, isto é, temos que recordar ou encontrar documentos/eventos/decisões/etc. similares, de forma a colocar a informação nova no enquadramento correcto de forma a sermos capazes de a processar de uma forma adequada. O cérebro humano é muito bom na realização destas associações, no entanto, é muito limitado no sentido de que o seu armazenamento contém um conjunto limitado de documentos amostra. Este conjunto limitado de amostras é limitado porque apenas são armazenadas no cérebro humano aquelas amostras que o utilizador ou o ser humano seja capaz de ler e de ouvir. No entanto, a real capacidade de ler ou de ouvir de um ser humano é quase nada em comparação com a quantidade de informação produzida em cada dia. Será, portanto, muito útil se uma máquina for capaz de armazenar amostras e então de as recuperar a partir das amostras assim armazenadas aquelas que possam ser vistas como uma "associação" para uma dada amostra de entrada. Um utilizador que seja provido com uma máquina como explicado

em ligação com a Figura 11 será, portanto, capaz de aceder de uma forma associativa a uma grande amostra de documentos que ele na realidade poderá não ver, ler, ou ouvir de todo, no entanto, que estão armazenados numa memória de base de dados 900 para permitir um acesso associativo como explicado. Se um utilizador então está interessado num certo aspecto de alguma peça de informação, ele pode criar o seu próprio ficheiro de entrada 910 e então a máquina recuperará para ele aquelas associações a partir do "conhecimento" pré-armazenado que sejam mais interessantes para ele.

De acordo com outro aspecto, um utilizador pode ser confrontado com uma nova entrada que ele seja incapaz de colocar no enquadramento correcto devido a ele não ter "conhecimento" acerca dele. Numa forma similar, um utilizador pode então fazer uso da máquina apresentada na Figura 11 para recuperar amostras a partir de um repositório ou base de dados 900 que possam ser vistas como "associações" do documento de entrada 910 desconhecido.

Um tal processo ou máquina como explicado em ligação com a Figura 11 é, portanto, de substancial assistência para qualquer ser humano que seja confrontado com o dilúvio de informação que ele encontra a cada dia. É útil quando realiza associações para colocar informação desconhecida no enquadramento correcto, de forma a recuperar amostras a partir de uma base de dados de conhecimentos que ele de outra forma poderia nunca

encontrar com base num documento auto-criado, ou de forma a classificar dados de entrada não classificados numa de um conjunto de classes de classificação pré-definidas.

Considerando que um ser humano quando processa informação que ele vai encontrando não faz nada mais do que realizar associações e levando a cabo classificações, uma tal máquina como explicado em ligação com a Figura 11 é verdadeiramente capaz de assistir o utilizador no processamento de informação e tomar conta de muito do trabalho que até agora tem que ser efectuado por um ser humano.

Lisboa, 28 de Setembro de 2012

REIVINDICAÇÕES

1. Método para recuperar a partir de um conjunto de documentos electrónicos aqueles documentos que sejam próximos em conteúdo de um documento electrónico utilizado como um documento de entrada, compreendendo o dito método:

A) Proporcionar um conjunto de documentos (D1,…, DN);

B) Gerar vectores de particularidades de cadeias de bits (110) para os documentos do referido conjunto, em que cada vector de particularidades representa uma representação dos referidos documentos e indica com cada uma das suas componentes binárias do vector a presença ou ausência de uma certa particularidade dentro do respectivo documento;

C) Formar uma matriz de bits de atributos (100; 800) a partir dos vectores de particularidades de cadeias de bits (110), consistindo a matriz de bits de atributos de bits individuais e representando cada bit um certo atributo para um certo documento pela indicação da presença ou ausência de uma certa particularidade dentro do documento respectivo;

D) Proporcionar um documento de entrada (200; 400; 700; 820);

E) Gerar uma pluralidade de fragmentos (S1, S2, S3; 410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) do referido documento de entrada (200; 400; 700; 820) que são partes do referido documento de entrada;

F) Para cada um dos referidos fragmentos, gerar um respectivo vector de particularidades de cadeia de bits de pesquisa (F1; F2; F3) representando o respectivo fragmento e indicando com cada uma das suas componentes de vector binárias a presença ou ausência de uma certa particularidade dentro do respectivo fragmento;

G) Efectuar uma pesquisa associativa para cada um dos referidos fragmentos utilizando a referida matriz de atributos de bits e o respectivo vector de particularidades de cadeia de bits de pesquisa (F1; F2; F3) representando o respectivo fragmento,

em que a pesquisa associativa para um fragmento respectivo é efectuada pela determinação para o fragmento e cada documento do referido conjunto de uma medida de similaridade individual entre o referido fragmento e o respectivo documento do referido conjunto,

em que a respectiva medida de similaridade individual é determinada efectuando uma operação lógica bit

a bit entre o vector de particularidades de cadeia de bits de pesquisa representando o respectivo fragmento e a cadeia de bits representando o documento respectivo na referida matriz de atributos, em que as referidas medidas de similaridade individuais, sendo cada uma determinada para um fragmento particular e um documento respectivo do referido conjunto e representando uma similaridade entre o referido um fragmento particular e o referido um documento respectivo do referido conjunto, são componentes de vector de um vector de medida de similaridade (SM1; SM2; SM3) associado ao respectivo fragmento particular e a todos os documentos do referido conjunto de documentos, em que para cada um dos referidos fragmentos é obtido um correspondente vector de medida de similaridade (SM1; SM2; SM3);

H) Para cada um dos referidos documentos do referido conjunto calcular uma respectiva medida de similaridade final individual a partir destas medidas de similaridade individuais, que de acordo com o passo G) são determinadas para o mesmo documento respectivo em relação aos referidos fragmentos, sendo as referidas medidas de similaridade final individuais componentes de vector de um vector de medida de similaridade final (FSM), que é obtida com isso com base nos referidos vectores de medida de similaridade (SM1, SM2, SM3);

I) Decidindo, com base nas referidas medidas de similaridade final individuais incluídas como componentes de vector no referido vector de medida de similaridade

final (FSM), quais documentos do referido conjunto podem ser vistos como tendo um conteúdo próximo do referido documento de entrada;

em que os referidos fragmentos (S1, S2, S3; 410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) são gerados pela segmentação do referido documento de entrada (200; 400; 700; 820) em segmentos, em que dos referidos segmentos o primeiro para o penúltimo segmentos têm uma dada dimensão e o último segmento tem a dimensão dada ou é mais curto do que a dimensão dada,

em que são gerados fragmentos de documento de entrada adicionais (450, 460, 470, 480; 720, 730) deslocando um ou mais fragmentos do referido documento de entrada para gerar fragmentos que se sobrepõem (410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) do referido documento de entrada (400; 700), para os quais são gerados respectivos vectores de particularidades de cadeia de bits de pesquisa e utilizados na referida pesquisa associativa.

2. Método da Reivindicação 1, em que a referida sobreposição entre os referidos fragmentos que se sobrepõem é uma sobreposição em unidades de um de:

caracteres;
sílabas;
fonemas;
palavras.

3. Método de uma das precedentes Reivindicações, em que o referido documento de entrada é obtido a partir da realização de um processo de reconhecimento de voz sobre um texto recitado por uma voz humana.

4. Método de uma das precedentes Reivindicações, em que cada um dos referidos documentos (D1, …, DN) do referido conjunto é classificado como pertencente a uma de uma pluralidade de classes (810), compreendendo o referido método ainda:

Realizar a referida pesquisa associativa, com base no número de documentos similares ao documento de entrada (820) que tenham sido encontrados nas classes individuais, classificando o referido documento de entrada como pertencendo a uma das referidas classes.

5. Método da Reivindicação 4, em que o número de classes é duas, uma para um conjunto relevante de documentos e uma para um conjunto não relevante de documentos;

Retornar os resultados de recuperação a partir da referida pesquisa associativa para um utilizador apenas se o referido documento de entrada tiver sido classificado como pertencente à referida classe relevante.

6. Método de uma das precedentes

Reivindicações, compreendendo ainda:

Optimizar os parâmetros do referido método tal como a dimensão dos fragmentos e número de elementos de sobreposição com base num dado esquema de classificação, compreendendo a referida optimização:

a) Utilizar documentos a partir do referido esquema de classificação para os quais a sua classificação seja já conhecida;

b) Classifica-los de acordo com a Reivindicação 6;

c) Variar os referidos parâmetros e repetir os passos a) e b);

Repetir o passo c) até que tenha sido encontrado um conjunto de parâmetros óptimo.

7. Método de uma das precedentes Reivindicações, em que

os dados de entrada de voz são representados através de fonemas para proporcionar o referido documento de entrada e os referidos documentos do referido conjunto, para os quais são gerados os referidos vectores de particularidades de cadeia de bits, contendo dados de fonemas;

ou

são utilizados dados de entrada de texto e os referidos documentos do referido conjunto, para os quais são gerados os referidos vectores de particularidades de cadeia de bits, contendo dados escritos em caracteres;

ou

os dados de entrada escritos são convertidos em dados de fonemas para proporcionar o referido documento de entrada e os referidos documentos do referido conjunto, para os quais são gerados os referidos vectores de particularidades de cadeia de bits, contendo dados de fonemas.

8. Método de acordo com uma das precedentes Reivindicações, em que de acordo com o passo H) a respectiva medida de similaridade final é calculada a partir das referidas medidas de similaridade individuais utilizando uma função de ponderação que atribui um peso mais elevado a medidas de similaridade individuais indicando uma maior similaridade entre o respectivo fragmento e um respectivo documento do referido conjunto e que atribui um peso mais baixo a medidas de similaridade individuais indicando uma similaridade mais baixa entre o respectivo fragmento e um respectivo documento do referido conjunto.

9. Método de acordo com uma das precedentes Reivindicações, em que a referida operação lógica bit a bit é uma operação lógica de conjunção “AND” bit a bit.

10. Método de acordo com a Reivindicação 9, em que as particularidades presentes são representadas por um "1" lógico e o número de "1"s lógicos resultantes a partir da operação lógica de conjunção "AND" bit a bit é utilizado como medida de similaridade individual.

Lisboa, 28 de Setembro de 2012

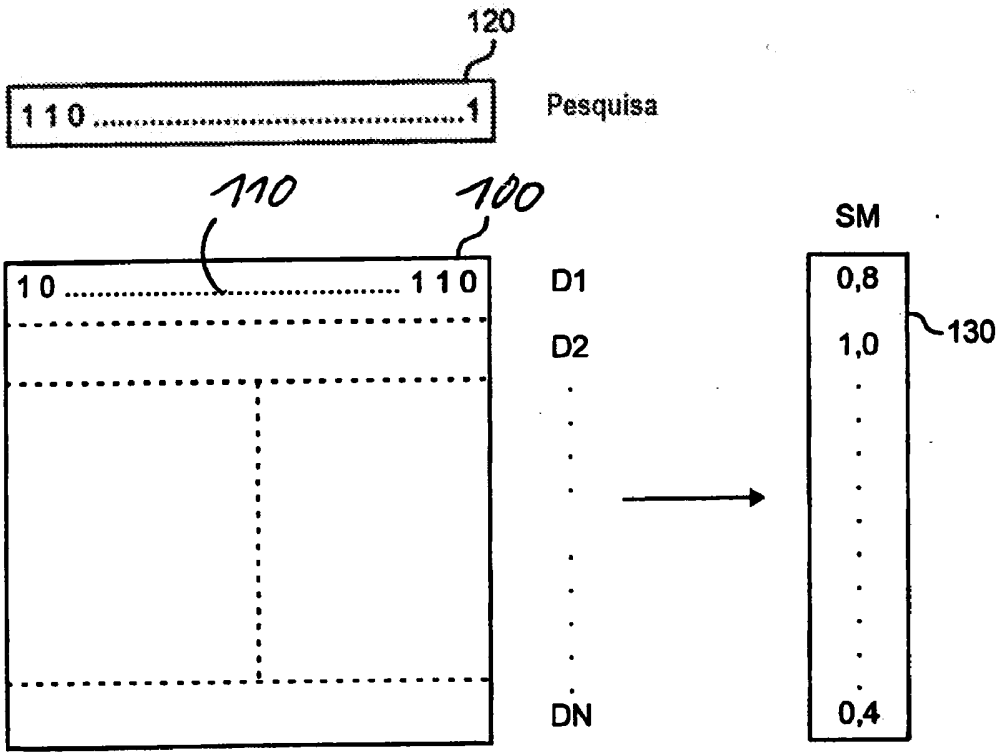


Figura 1

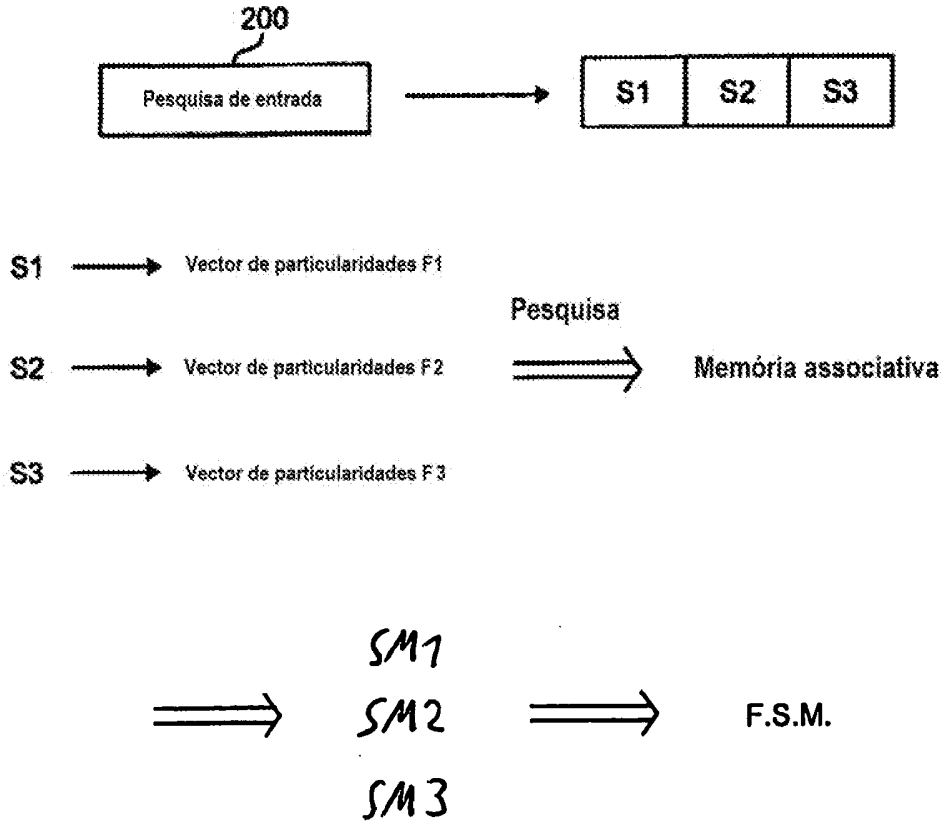


Figura 2

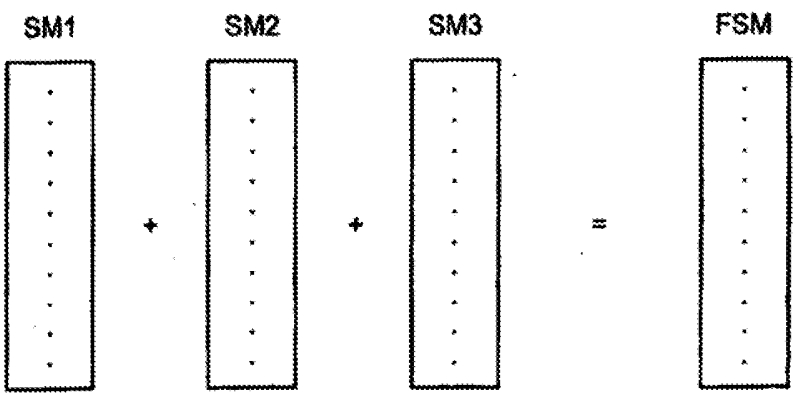


Figura 3

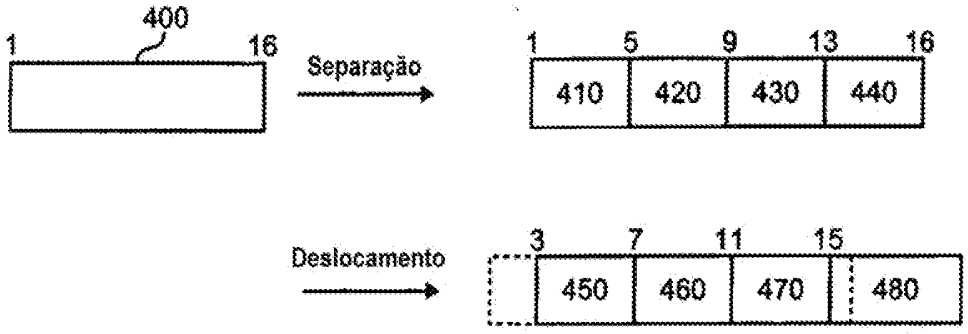


Figura 4

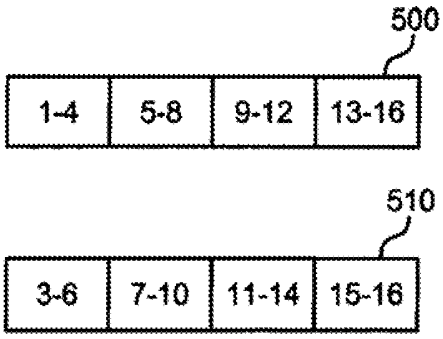


Figura 5

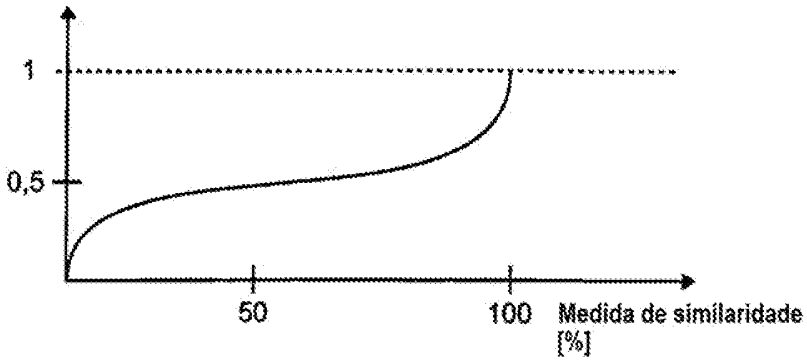


Figura 6

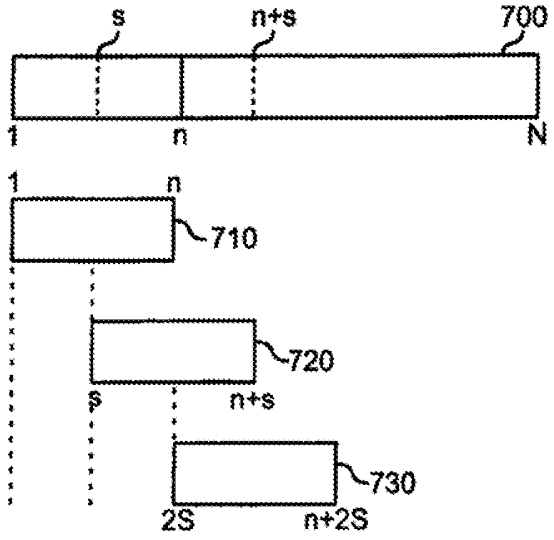


Figura 7

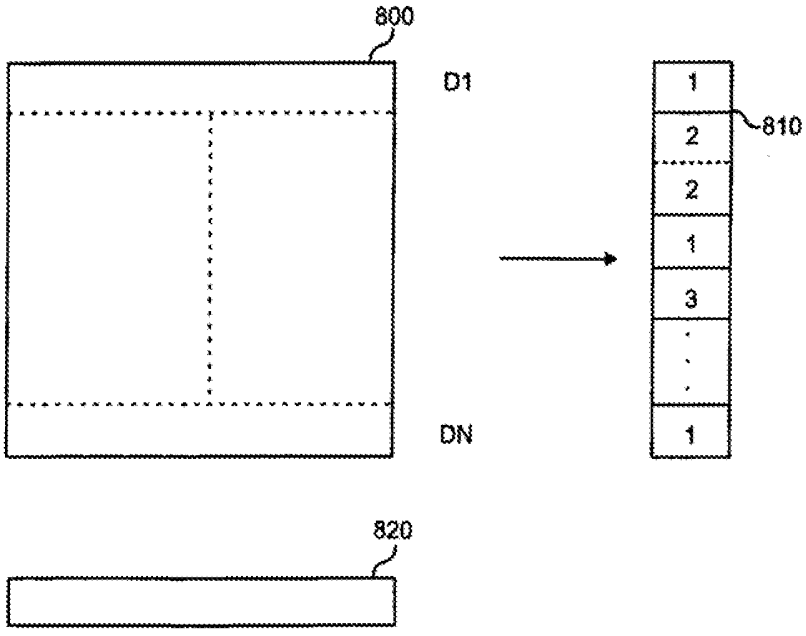


Figura 8

Estatísticas de recuperação

Classe	1	2	3
N.º de Concordâncias	5	85	12

Figura 9

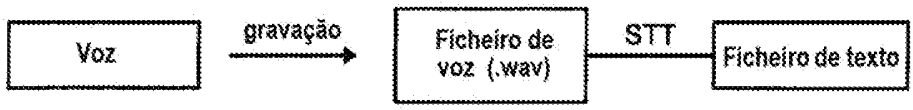


Figura 10

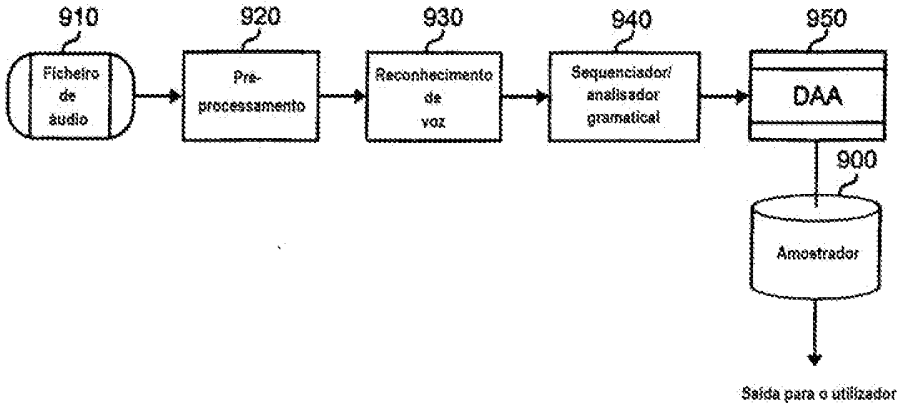

910
Ficheiro de audio
920
Pre-processamento
930
Reconhecimento de voz
940
Sequenciador/ analisador gramatical
950
DAA
900
Amostrador
Saída para o utilizador

Figura 11

REFERÊNCIAS CITADAS NA DESCRIÇÃO

Esta lista de referências citadas pelo requerente é apenas para conveniência do leitor. A mesma não faz parte do documento da patente Europeia. Ainda que tenha sido tomado o devido cuidado ao compilar as referências, podem não estar excluídos erros ou omissões e o IEP declina quaisquer responsabilidades a esse respeito.

Documentos de patentes citadas na descrição

- US 5778362 A
- EP 0750266 A1
- US 5918223 A
- EP 1049030 A1

Literatura que não é de patentes citada na descrição

- **Berkovich, S. ; El-Qawasmeh, E. ; Lapir, G.** Organization of near matching in bit attribute matrix applied to associative access methods in information retrieval. *Proceedings of the 16th IASTED International Conference Applied Informatics,* 23 February 1998
- **Lapir, G. M.** Use of associative access method for information retrieval systems. *PROC. 23rd PITTSBURGH CONF. ON MODELING AND SIMULATION,* 1992, 951-958
- Syllable-based Chinese text/spoken document retrieval using text/speech queries. **BO-REN BAI et al.** INTERNATIONAL JOURNAL OF PATTERN RECOGITION AND ARTIFICAL INTELLIGENCE. WORLD SCIENTIFIC, SINGAPORE, August 2000, vol. 14, 603-616