República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) **BR 10 2013 016668-5 A2**

* B R 1 0 2 0 1 3 0 1 6 6 6 8 A 2 *

(22) Data de Depósito: 27/06/2013

(43) Data da Publicação: 04/08/2015
(RPI 2326)

(54) Título: SISTEMA E MÉTODO PARA BUSCA FONÉTICA DE DADOS

(51) Int.Cl.: G06F17/30; G06F17/27

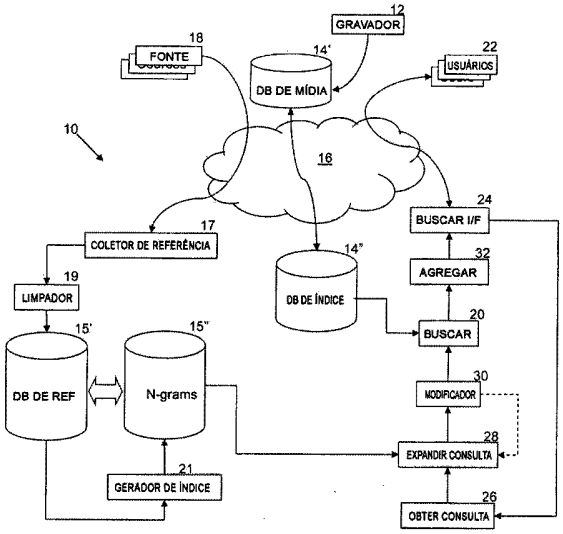
(52) CPC: G06F17/30746; G06F17/27

(30) Prioridade Unionista: 06/09/2012 US 13/605,084

(73) Titular(es): AVAYA INC.

(72) Inventor(es): BRIAN ANDREW MELLOR , GARETH ALAN WYNN, KEITH MICHAEL PONTING, MALCOLM FINTAN WILKINS

(57) Resumo: SISTEMA E MÉTODO PABA BUSCA FONÉTICA DE DADOS. Um método para foneticamente buscar mídia incluindo uma pluralidade de trilhas de áudio é revelado onde cada trilha de áudio é indexada para fornecer uma representação fonética da trilha de áudio. O método compreende obter uma consulta de busca de texto e buscar a consulta de texto contra um conjunto de documentos de referência para obter um subconjunto de documentos pseudo-relevantes. Os documentos pseudo-relevantes são examinados para um conjunto de expressões de busca caracterizando os documentos pseudo-relevantes. Uma representação fonética correspondendo a pelo menos algumas do conjunto de expressões de busca é fornecida e para cada das representações fonéticas das expressões de busca, as representações fonéticas indexadas para uma ou mais da pluralidade de trilhas de áudio são foneticamente buscadas para fornecer quaisquer indicadores da incidência da expressão de busca em uma ou mais trilhas de áudio. Os indicadores resultantes da busca fonética são combinadas em um conjunto de resultados combinados para cada do conjunto de expressões de busca; e os resultados combinados retornados.



SISTEMA E MÉTODO PARA BUSCA FONÉTICA DE DADOS

PEDIDO RELACIONADO

O presente pedido se refere ao pedido US no. 13/605.055 depositado em 6 de setembro de 2012 e intitulado "A system and method for phonetic searching of data" (ref.: 512125-US-NP/P105534us00/A180FC) e que é incorporado aqui a título de referência.

Campo da invenção

A presente invenção refere-se a um sistema e método para busca fonética de dados.

Antecedentes

A expansão de consulta para busca de texto é conhecida. Em algumas aplicações de expansão de consulta, uma expressão de busca de texto é comparada com um conjunto de documentos de referência para selecionar a partir desses documentos um conjunto relativamente pequeno de expressões relevantes para a expressão de busca. Esse conjunto de expressões é então utilizado para buscar um conjunto alvo de documentos. Os resultados para cada busca a partir do conjunto expandido de expressões são combinados para fornecer um resultado final de busca - por exemplo, por classificar os documentos mais relevantes a partir do conjunto de resultados e remover resultados em duplicata.

Fatores importantes em expansão de consulta para busca de texto são "de origem" e "para remoção de palavra", por exemplo, como revelado por Pierre Jourlin, Sue E. Johnson, Karen Sparck Jones, e Philip C. Woodland. "General query expansion techniques for spoken document retrieval," páginas 8-13 em Proceedings ESCA Tutorial and Research Workshop "Accessing information in spoken audio,"

Cambridge, UK, abril de 1999. Desse modo, por exemplo, certos termos de consulta são reduzidos a sua forma de raiz e palavras comuns são removidas da expressão de busca. Tipicamente, o corpus de documentos de referência é reduzido a "sacos de palavras", registrando:

. Para cada palavra, ou termo, "freqüência de documento", que é o número de documentos distintos nos quais esse termo ocorre;

. Para cada documento e termo, "freqüência de termo" que é o número de vezes que o termo ocorre naquele documento.

Essa abordagem, entretanto, não seria apropriada para busca fonética de um banco de dados de mídia incluindo trilhas de áudio porque utilizar uma expressão de busca resumida ou de origem poderia produzir um equivalente fonético improvável de ser encontrado em fala normal gravada no banco de dados de mídia. Igualmente, dividir os documentos de referência em seus fonemas mais comuns ou distintos seria sem sentido.

Deve ser também observado que bancos de dados de áudio não têm tipicamente ou necessariamente um banco de dados de texto correspondente (um motivo sendo que transcrição de texto é extremamente intenso em processador) e se o banco de dados tivesse sido transcrito em texto, seria muito mais fácil buscar o banco de dados de texto para encontrar uma entrada correspondente no banco de dados de áudio. Desse modo, a necessidade de busca fonética seria evitada.

US 2010/0211569, Avaya, revela um sistema que utiliza dados de treinamento que compreendem uma

pluralidade de documentos de treinamento. Cada da pluralidade de documentos de treinamento compreende um token(s) de treinamento. A pluralidade de documentos de treinamento é agrupada em uma pluralidade de clusters baseados pelo menos em um token de treinamento na pluralidade de documentos de treinamento. Cada cluster contém pelo menos um documento de treinamento. Uma (s) consulta(s) Booleana é/são gerada(s) para um cluster com base em uma ocorrência de pelo menos um token de treinamento em um documento de treinamento na pluralidade de documentos de treinamento. O sistema obtém dados de produção que compreendem uma pluralidade de documentos de produção. Cada da pluralidade de documentos de produção compreende um token(s) de produção. A(s) consulta booleana(s) é/são então executadas nos dados de produção.

No campo de busca fonética, US2009/0326947 revela o uso de um mecanismo de categorização de tópico, porém com base em torno de treinamento explícito com material de áudio rotulado de acordo com uma hierarquia de tópico pré-especificada.

De modo semelhante, Timothy J. Hazen, Fred Richardson e Anna Margolis, "Topic identification from áudio recordings using Word and phone recognition lattices," Proceedings of the IEEE Workshop on automatic speech recognition and understanding, Kyoto, Japão, dezembro de 2007; e Christophe Cerisara, "Automatic discovery of topics and acoustic morphemes from speech", Computer speech and language, v. 23, no. 2, pág. 220-239, abril de 2009 - ambos iniciam de dados de treinamento rotulados com uma lista preestabelecida de tópicos e se

referem a determinação de seqüências de fonema relacionadas a tópico e fragmentos de palavras.

É um objetivo da presente invenção fornecer busca aperfeiçoada de bancos de dados de áudio.

Sumario

A presente invenção compreende um método para busca fonética de dados de acordo com a reivindicação 1.

Em um aspecto adicional, é fornecido um produto de programa de computador armazenado em um meio de armazenagem legível em computador que quando executado em um processador é organizado para realizar as etapas de qualquer uma das reivindicações 1 a 23.

Ainda em um aspecto adicional, é fornecido um sistema de busca fonética organizado para executar as etapas de qualquer uma das reivindicações 1 a 23.

A presente invenção permite que um usuário procure a ocorrência de tópicos em material de áudio, onde os tópicos são especificados por uma string de busca, porém com o desejo de ampliar a busca além de ocorrências puramente contendo as palavras na string de busca.

Breve descrição dos desenhos

Uma modalidade da invenção será descrita agora, como exemplo, com referência ao desenho em anexo, no qual:

A figura 1 mostra esquematicamente as etapas envolvidas em busca fonética de acordo com uma modalidade da presente invenção.

Descrição da modalidade preferida

Com referência agora à figura 1, que mostra um sistema de busca fonética 10 de acordo com uma modalidade da presente invenção. Um sistema de gravação 12 provê um

banco de dados 14' de arquivos de mídia incluindo trilhas de informações de áudio que devem ser buscadas. A mídia pode compreender, por exemplo, televisão broadcast ou programas de rádio ou em outras implementações, a mídia pode compreender gravações de contatos de um centro de contato (não mostrado) entre usuários e agentes do centro de contato, ou ainda em outras implementações a mídia pode compreender gravações de chamadas de vídeo; ou eventos gravados em vídeo. Tipicamente, o acesso aos arquivos de mídia é fornecido através de uma rede 16 que pode ser qualquer de uma LAN, WAN ou Internet. Dependendo das exigências e recursos, os arquivos de mídia podem ser copiados de modo que são localmente disponíveis para o sistema de busca 10.

Informações fonéticas são extraídas para cada arquivo de mídia e essas são armazenadas em um banco de dados de índice 14" com informações de índice no banco de dados 14" apontando para informações de áudio correspondentes no banco de dados 14'. Um esquema particularmente útil para implementar essa indexação é descrita no pedido US no. 13/605.055 intitulado "A system and method for phonetic searching of data" (Ref: 512125-US-NP / P105534us00 / A180FC) e que é incorporado aqui a título de referência.

Na modalidade, informações fonéticas extraídas dos arquivos de áudio são mostradas armazenadas localmente no banco de dados de índice 14". Entretanto, em outras implementações, um motor de busca fonética 20 e o banco de dados de índice 14" podem ser remotos do restante do sistema 10 com uma interface de busca solicitando ao motor

de busca fonética 20 para fazer buscas específicas como necessário. Em qualquer caso, pelo menos informações fonéticas correspondendo às informações de áudio a serem buscadas necessitam ser disponíveis para o motor de busca fonética 20.

Separadamente, o material de fonte 18 para um banco de dados de referência 15' é gerado por um coletor 17. Idealmente, o material para esse banco de dados 15' compreende uma coleção de material de texto geral, com o máximo possível cada arquivo de banco de dados ou objeto contendo texto relevante para um ou um número pequeno de tópicos relacionados, esses tópicos por sua vez sendo de interesse para usuários e referentes ao assunto das trilhas de áudio armazenadas no banco de dados 14'.

O material de fonte 18 poderia incluir web sites de broadcaster que freqüentemente incluem artigos de notícias correspondendo a material de programa broadcast - cada artigo ou seção substancial de um artigo representando um objeto/documento de referência separado no banco de dados 15'.

Em um caso específico onde os arquivos de mídia compreendem broadcasts parlamentares, o material de fonte 18 pode compreender transcrições de tais broadcasts que são normalmente disponíveis separadamente.

Outras fontes úteis 18 poderiam ser manuais de usuário para produtos sendo tratados por agentes de um centro de contato. Esses poderiam ser divididos por seção para fornecer arquivos/objetos de banco de dados de referência separados referentes a tópicos dados.

Não obstante, fontes 18 poderiam ser mais gerais

e poderiam compreender, por exemplo, feeds de fontes de ligação em rede social como Twitter ou Facebook.

Utilizando um número limitado de fontes como os exemplos acima permite que material seja amplamente automaticamente limpo e dividido em objetos separados no banco de dados. Assim, por exemplo, o layout de um web site de broadcaster serão relativamente consistente e similarmente manuais de produto e outra literatura de um dado vendedor que fornece um centro de contato devem ser razoavelmente consistentes. Isso permite que material não útil, por exemplo, cabeçalhos repetidos através de todos os artigos/seções sejam retirados quando é coletado pelo coletor 17 ou subseqüentemente por um limpador 19 como descrito abaixo.

Não obstante, deve ser reconhecido que a invenção não é limitada à coleta de qualquer número ou qualquer forma específica de material de fonte.

Material coletado pode ser limpo quando recebido por um limpador 19 antes de ser gravado no banco de dados; e/ou, além disso, ou alternativamente o material de banco de dados pode ser limpo após ser gravado no banco de dados 15'. O banco de dados de referência 15' pode ser continuamente atualizado pelo coletor 17 e limpador 19 e, por exemplo, após ter atingido a capacidade, documentos mais antigos ou material redundante podem ser removidos.

Como mencionado acima, pode ser útil limpar o banco de dados de referência até certo ponto para assegurar que os dados mais úteis são retidos para expandir uma consulta de busca. Alguns exemplos da limpeza de documentos de referência incluem:

a) Transliteração - assegurar que todas as seqüências de caractere estejam compreendidas em uma codificação especificada, por exemplo, ASCII ou Unicode. Isso é porque finalmente expressões de busca que são escolhidas necessitarão ser convertidas em um fluxo fonético e há pouca vantagem em reter qualquer material no banco de dados de referência que não seja prontamente convertível em formato fonético.

b) Substituição de seqüências que parecem ser fórmulas matemáticas com, por exemplo, "+++".

c) Substituição de seqüências UTF-8 em páginas de rede não codificadas UTF-8 com caracteres equivalentes.

d) Tradução de caracteres com um oitavo conjunto de bit para um caractere ASCII equivalente ou entidade HTML - por exemplo, o hexadecimal 90 é traduzido em "’" (aspas únicas direita).

e) Substituição de números ou datas com seqüências genéricas, por exemplo, "xNx" para números e "yDy" para datas. Isso evita a ocorrência de datas ou números em expressões de busca puramente porque (em um banco de dados relativamente escasso) associações espúrias podem aparecer entre números e tópicos. Não obstante é reconhecido que essa abordagem tem a desvantagem de ignorar quaisquer associações semânticas de seqüências

de data ou número específicas, como (no RU) 1066 sendo a data de uma invasão ou (nos EUA) 4 de julho. Em bancos de dados maiores, poderia haver resistência estatística suficiente para tornar essa substituição desnecessária.

f) Tradução de outros caracteres não ASCII em um ASCII apropriado quase equivalente ou em "~" se não houver equivalente óbvio.

g) Alguns documentos de fonte incluindo, por exemplo, páginas de rede, poderiam compreender uma estrutura de árvore com cada nó da árvore compreendendo fragmentos de texto de documento. Em algumas implementações, um nó e seu texto poderiam ser retidos se e somente se: não tiver hiperlinks e mais de um número mínimo especificado (digamos 10) de palavras ou mais do que outro limite (digamos 20) de palavras por hiperlink.

h) Novamente, onde o material de fonte é tirado de um web site, documentos de página individuais podem conter certos nós mencionados que são conhecidos como contendo freqüentemente parágrafos diretamente abaixo daquele nó duplicado em outros documentos. Quaisquer tais nós mencionados (exceto um documento de nível superior) poderiam ser descartados.

i) Também nós em documentos estruturados que compreendem certas palavras chave, por

exemplo, "renúncia" são genericamente conhecidos como compreendendo chapas grossas e tais nós podem ser descartados de documentos armazenados no banco de dados 15'.

j) Outras técnicas para identificar chapas grossas são descritas em Christian Kohlschutter, Peter Fankhauser, Wolfgang Nejdl: "Boilerplate detection using shallow text features". WSDM 2010: 441-450 e esses podem ser também implementados em certas modalidades da presente invenção.

k) Remoção de documentos e/ou parágrafos em duplicata - segunda e subseqüente ocorrências de quaisquer documentos/parágrafos podem ser removidos com base em uma "soma de verificação", por exemplo, gerada com MD4, computada para cada documento/parágrafo após todas as etapas de limpeza acima. (Essa etapa é importante para evitar expansão de consulta prestando atenção em demasia para termos que aparecem em duplicatas).

l) Parágrafos que ocorrem freqüentemente (por exemplo, mais de 500) também podem ser descartados de documentos no banco de dados de referência 15'.

Após material de referência ter sido limpo, um conjunto de expressões (n-gramas) é gerado para cada objeto/documento separado do banco de dados de referência 15' por um gerador de índice 21. Ao passo que para busca de texto, um documento poderia ser dividido em um saco de

palavras com uma contagem mantida de cada ocorrência de palavra (sem parar) no documento, no presente caso, um documento/objeto é associado a conjuntos de N-gramas, cada N-grama compreendendo uma seqüência de N palavras a partir do documento de referência, com N tipicamente variando de 2-5. Desse modo, uma contagem é mantida de cada instância de par de palavras, triplo de palavras, quad. Etc., que aparece em documentos respectivos do banco de dados de referência 15'.

Para racionar o número de N-gramas mantido para qualquer documento/objeto dado e melhorar a relevância de sua contagem, algumas das seguintes etapas podem ser tomadas para igualar instâncias separadas de N-gramas para fins de contagem:

• o texto de uma primeira ocorrência de um N-grama é gravado como uma forma de referência do N-grama, porém uma forma "extraída" é utilizada para comparar e contar em que:

° caso é ignorado (porém uma instância de letra minúscula substituiria uma instância de letra maiúscula como a forma de referência);

° trailing ' é removido, de modo que, por exemplo, Sábados' e sábado são contados como equivalentes;

° trailing 's é removido, de modo que por exemplo, Sábado's e sábado são contados como equivalentes;

° todos os caracteres não alfabéticos são removidos, o que pode levar a alguma ambigüidade, porém permite, por exemplo, que painkilling, pain-killing e pain killing sejam tratados como equivalentes;

° incorporado ' e 's são removidos, de modo que,

por exemplo, "BBC's correspondent" e "BBC correspondent" são contados como equivalentes.

• palavras de parar são aparadas de qualquer extremidade de um N-grama candidato para fins de comparar com outros N-gramas e contagem;

• N-gramas são somente contados se atenderem as seguintes limitações heurísticas:

° o número de palavras distintas N deve ser entre 2 e 5.

° o comprimento fonético deve ser pelo menos 12 fonemas na pronúncia mais curta.

° o número mínimo de ocorrências no conjunto de documentos de referência é ajustado como 2.

• N-gramas não podem limitar caracteres ou seqüências como "~", "+++", "x**N**x" ou "yD" que foram inseridos no estágio de limpeza.

• se dois M-gramas (M<N) obtidos por remover a palavra dianteira e quaisquer palavras de parada adjacentes à mesma e por remover a palavra traseira e quaisquer palavras de parar adjacentes atendem ambas as limitações heurísticas acima e, portanto seriam incluídas, a instância posterior de N-grama não é contada separadamente. O resultado disso é um conjunto de frases de busca candidatas indexadas 15" associadas a cada objeto/documento limpo do banco de dados de referência 15'.

Como o banco de dados alvo 14" compreende fluxos fonéticos correspondendo a frases faladas, somente as formas mais limitadas de origem das frases de busca candidatas são empregadas pelo gerador de índice 21 - assim, por exemplo, somente certas palavras de parar

poderiam ser aparadas de qualquer extremidade da string de busca.

Outro processamento das frases de busca candidatas poderia incluir processamento de linguagem natural (NLP) das seqüências de palavra para converter formas gravadas em uma ou mais strings alternativas lembrando mais estreitamente a fala normal. Por exemplo, a string "2012" poderia ser convertida em "vinte doze" se o contexto sugerisse uma data. Múltiplas alternativas originam se o contexto for ambíguo ou houver formas faladas variantes - "dois mil e doze" seria outro modo de falar o ano em um contexto de data. O processo relacionado de traduzir de múltiplas formas faladas possíveis em uma forma gravada consistente é conhecido como "normalização de texto inversa" (vide, por exemplo, o pedido de patente US 2009/0157385).

Após o índice de frases de busca 15" ser fornecido, pode ser agora tornado disponível para expansão de consulta.

Na presente modalidade, os usuários 22 acessam o sistema de busca 10 através de uma interface de busca 24. Tipicamente isso poderia compreender um aplicativo de rede acessado através da rede 16, não obstante, o aplicativo poderia ser igualmente implementado como um aplicador de servidor-cliente dedicado ou independente.

Os usuários entram sua consulta de busca compreendendo uma string de texto. A busca de áudio fonético funciona melhor em expressões de busca mais longas, e assim o objetivo de um expansor de consulta 28 é encontrar conjuntos de seqüências de palavras (N-gramas)

como frases de busca possíveis com base na string de busca de texto inicial fornecida através da interface de busca.

Na presente modalidade, o expansor de consulta 28 opera em 2 fases:

- em uma primeira fase, um motor de busca de texto do tipo convencional, por exemplo, Lucene, é empregado para localizar uma seqüência ordenada de documentos (pseudo-relevantes) a partir do banco de dados de referência 15' que considera relevantes para a consulta de busca inicial. Na modalidade, cada documento pseudo-relevante recebe uma ponderação de relevância associada e qualquer esquema pode ser empregado, por exemplo, BM25 descrito em: Stephen Robertson e Hugo Zaragoza. SIGIR 2007 Tutorial 2d "The probabilistic relevance model: Bm25 and beyond" em Wessel Kraaij, Arjen P. de Vries, Charles L.A. Clarke, Norbert Fuhr, e Noriko Kano, editores, SIGIR ACM, 2007, para pesar os documentos. Em uma implementação, o número de documentos pseudo-relevantes é definido em 50. Alguns desses documentos, evidentemente, podem não ser relevantes (ou tão relevantes quanto aparecem para o motor de busca) e opcionalmente a interface de busca 24 poderia ser organizada para permitir que o usuário 22 examine os documentos pseudo-relevantes retornados e aceite/rejeite algum número dos documentos.

- em uma segunda fase, um número, tipicamente 20, de frases de busca é escolhido do conjunto de N-gramas candidatos associados ao conjunto de documentos pseudo-relevantes e ordenados por relevância. A marcação para cada N-grama se baseia nas estatísticas de ocorrências dos N-gramas nos documentos pseudo-relevantes produzidos pelo

motor de busca; a ponderação de relevância de documento produzida pela operação de primeira fase do motor de busca; e possivelmente outras estatísticas pertinentes ao banco de dados de referência 15', 15" como um todo, por exemplo, uma capacidade de distinção de N-grama no banco de dados de referência como um todo em vez de apenas no conjunto de documentos pseudo-relevantes.

O conjunto resultante de expressões de busca fornecido pelo expansor 28 pode, por sua vez, ser fornecido para um modificador 30 antes que a busca seja executada. Assim, em uma implementação, o conjunto de expressões de busca é apresentado através da interface de busca 24 (conexão não mostrada) para o usuário 22 para verificação manual, aumento e/ou deleção. Seria também possível para o modificador 30 retornar as expressões especificadas pelo usuário (ou verificadas) para o expansor 28 para repetir o processo de expansão de consulta com base em expressões modificadas para refinar ou estender o conjunto de termos.

Em outras implementações, o modificador 30 poderia utilizar os métodos revelados em Koen Deschacht e outros 2012, "The latent words language model", Computer speech and Language, 26, 384-409 para expandir o conjunto de expressões de busca para incluir sinônimos e/ou encontrar mais frases/palavras relacionadas.

Após o conjunto expandido de expressões de busca ter sido finalmente determinado, é submetido a um motor de busca 20 que utiliza uma representação fonética de cada do conjunto de expressões de busca para buscar representações fonéticas de informações de áudio armazenadas no banco de dados de índice 14". Em uma implementação, o motor de busca

fonética de meio de extração de áudio Aurix varre o banco de dados de índice 14" em relação a ocorrências de cada do conjunto de expressões de busca e retorna um fluxo de hits de busca, cada incluindo: uma identidade do arquivo de mídia no banco de dados 14' onde a expressão de busca ocorre, informações de tempo indicando a localização no arquivo de mídia da expressão de busca, identidade da expressão de busca e possivelmente uma marcação de casamento. Em uma implementação particularmente vantajosa da presente invenção, o motor de busca 20 e banco de dados de índice 14" são implementados em uma plataforma de partilha de arquivo distribuído (DFS) como revelado no pedido US no. 13/605.055 intitulado "A system and method for phonetic searching of data" (Ref.: 512125-US-NP / P105534us00/A180FC) e que é incorporado aqui a título de referência. Aqui informações de áudio do banco de dados 14' são indexadas em um conjunto de pastas de arquivo 14" tornando o desempenho de tarefas de busca de processamento paralelo bem eficientes.

Em qualquer caso, o motor de busca 20 provê o fluxo de hits de busca à medida que são gerados para cada expressão de busca a um mecanismo de agregação 32 que processa os hits. O agregador 32 pode executar qualquer combinação das seguintes etapas:

a) Limitar, com base em marcações de casamento, para remover os hits menos relevantes;

b) Executar remoção de sobreposição onde um hit é removido se outro, marcação melhor, hit sobrepor o mesmo em mais de uma fração especificada (digamos 30%) da duração do mais

curto dos dois hits;

c.1) contar as ocorrências de hits de busca de modo que um hit seja somente relatado se, em uma janela de tempo específica (default 10 segundos), pelo menos uma contagem mínima dada (digamos 2) de hits para expressões de busca distintas no conjunto expandido de expressões de busca forem encontrados;

c.2) alternativamente, em vez de exigir um número mínimo de casamentos para expressões distintas, casamentos para quaisquer das expressões podem ser contadas assim, por exemplo, para um conjunto de expressões de busca "A", "B" e "C" onde dois casamentos eram exigidos, então dois casamentos para "A" poderiam ser suficientes para acionar um hit;

d) executar uma soma ponderada similar àquela revelada em J. Wright, M. Carey, E. Parris, "Improved topic spotting through statistical modelling of keyword dependencies, em Proc. IEEE ICASSP, vol. 1, IEEE, Detroit, 1995, pág. 313-316), exceto que os pesos, em vez de serem treinados de material de áudio rotulado, são derivados de (ou uma combinação de): (i) as marcações de expressão de busca e quaisquer estatísticas obtidas durante expansão de consulta e (ii) a marcação de casamento de busca fonética correspondendo ao hit de busca específico.

Em algumas implementações, no estágio

modificador, a interface de busca 24 que permite ao usuário 22 ajustar o conjunto expandido de expressões de busca poderia ser organizada para permitir que o usuário especifique combinações booleanas das expressões de busca no conjunto expandido de expressões de busca. Desse modo, os resultados do motor de busca 20 poderiam ser combinados pelo agregador 32 de acordo com a lógica booleana especificada para as expressões de busca.

Em qualquer caso, após a agregação ser concluída ou realmente mesmo quando hits estão sendo gerados, o conjunto de resultados de busca é passado de volta para o usuário 22 através da interface de busca 24.

Há, evidentemente, muitas possibilidades para estender a funcionalidade da modalidade descrita acima. Por exemplo, resultados de busca não têm de ser passados de volta para o mesmo usuário que formulou a consulta original; nem uma consulta tem de ser formulada do princípio cada vez que uma busca é executada. Por exemplo, será visto que a consulta final que é utilizada pelo motor de busca 20 para fornecer o que poderia ser uma análise de mídia bem útil poderia ser salva e rotulada, por exemplo, com um identificador de tópico. A seguir, a consulta salva poderia ser repetida pelo usuário original posteriormente, talvez limitada para a mídia adquirida mais recentemente atendendo a consulta; ou alternativamente a consulta poderia ser executada novamente imediatamente por quaisquer usuários que tenham interesse no tópico identificado pelo rótulo de busca salvo. Realmente, resultados de consulta podem ser proativamente disseminados através de redes sociais de indivíduos que indicaram um interesse no

identificador de tópico na forma de newsfeeds.

Desse modo, será reconhecido que para fins de simplicidade, na modalidade ilustrada acima, mídia é mostrada como sendo armazenada em um banco de dados 14'. Entretanto, as informações de mídia sendo buscadas poderiam ser igualmente informações de mídia streamed, ao vivo sendo indexadas e varridas com consultas de busca expandida para detectar automaticamente tópicos sendo broadcast e notificar usuários interessados da ocorrência de um tópico de interesse em um programa sendo broadcast.

A invenção não é limitada à(s) modalidade(s) descrita(s) aqui, porém pode ser emendada ou modificada sem se afastar do escopo da presente invenção.

REIVINDICAÇÕES

1. Método para buscar foneticamente mídia, a mídia incluindo uma pluralidade de trilhas de áudio, o método sendo caracterizado pelo fato de que compreende as etapas de:

a) Indexar cada trilha de áudio para fornecer uma representação fonética de cada trilha de áudio;

b) Obter uma consulta de busca de texto;

c) Buscar a consulta de texto contra um conjunto de documentos de referência para obter um subconjunto de documentos pseudo-relevantes;

d) Examinar os documentos pseudo-relevantes para um conjunto de expressões de busca caracterizado os documentos pseudo-relevantes;

e) Fornecer uma representação fonética correspondendo a pelo menos algumas do conjunto de expressões de busca;

f) Para cada das representações fonéticas das expressões de busca, buscar foneticamente as representações fonética indexadas para uma ou mais da pluralidade de trilhas de áudio para fornecer quaisquer indicadores da incidência da expressão de busca em uma ou mais trilhas de áudio;

g) Combinar os indicadores resultantes da busca fonética em um conjunto de resultados combinados para cada do conjunto de expressões de busca; e

h) Retornar os resultados combinados.

2. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que compreende armazenar a mídia em um banco de dados remoto e fornecer as representações fonéticas das trilhas de áudio localmente.

3. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que compreende extrair documentos de referência a partir de qualquer combinação de: web sites, manuais de produto, fontes de ligação em rede social ou news feeds.

4. Método, de acordo com a reivindicação 3, caracterizado pelo fato de que compreende processar documentos de referência extraídos de acordo com qualquer combinação das seguintes regras:

• substituir números ou datas específicas nos documentos de referência com strings genéricas;

• substituir fórmulas nos documentos de referência com strings genéricas;

• remover chapas grossas dos documentos de referência;

• substituir caracteres não padrão nos documentos de referência com strings genéricas;

• em documentos estruturados compreendendo nós com fragmentos de texto, remover nós não distintos conhecidos;

• remover documentos de referência duplicados e parágrafos duplicados através de documentos de referência; e

• remover parágrafos que ocorrem freqüentemente a partir de documentos de referência.

5. Método, de acordo com a reivindicação 1,

caracterizado pelo fato de que compreende para cada documento de referência, gerar conjuntos de expressões compreendendo N palavras e contar instâncias de cada expressão em cada documento de referência.

6. Método, de acordo com a reivindicação 5, caracterizado pelo fato de que a contagem compreende: somar contas para expressões que somente diferem de acordo com qualquer combinação de: caso, apóstrofos, plurais, hifenação, palavras de parar traseira e avançada; descontar expressões com um comprimento fonético menor do que um limite; ou descontar expressões que parecem menos do que um número limite de vezes no conjunto de documentos de referência.

7. Método, de acordo com a reivindicação 5, caracterizado pelo fato de que compreende remover palavras de parar avançada e traseira de pelo menos algumas das expressões.

8. Método, de acordo com a reivindicação 5, caracterizado pelo fato de que compreende ainda fornecer uma ou mais formas faladas alternativas correspondendo a pelo menos algumas das expressões.

9. Método, de acordo com a reivindicação 5, caracterizado pelo fato de que a etapa c) compreende fornecer uma lista classificada de documentos pseudo-relevantes de acordo com sua relevância para a consulta de busca.

10. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que a etapa c) compreende a etapa de: responsivo à interação de usuário, ajustar o conjunto de documentos pseudo-relevantes.

11. Método, de acordo com a reivindicação 9, caracterizado pelo fato de que a etapa d) compreende escolher o conjunto de expressões de busca pelo menos como uma função da classificação dos documentos pseudo-relevantes nos quais as expressões de busca ocorrem.

12. Método, de acordo com a reivindicação 11, caracterizado pelo fato de que a etapa d)compreende escolher o conjunto de expressões de busca pelo menos como uma função da contagem das expressões de busca nos documentos pseudo-relevantes nos quais as expressões de busca ocorrem.

13. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que compreende ainda: antes da etapa e) e responsivo à interação de usuário, ajustar o conjunto de expressões de busca.

14. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que compreende repetir as etapas c) e d) com cada do conjunto de expressões de busca e fundir o conjunto resultante.

15. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que a combinação compreende: remover sobreposições nas trilhas de áudio a partir dos resultados de busca; ou fornecer combinações Booleanas especificadas para o usuário dos resultados de busca.

16. Método, de acordo com a reivindicação 2, caracterizado pelo fato de que o banco de dados de mídia compreende: gravações de contatos processados por um centro de contato; um de programas de broadcast de rádio ou televisão; gravações de chamadas de vídeo; ou eventos gravados de vídeo.

17. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que a mídia compreende mídia de broadcast ao vivo, chamadas de vídeo ou áudio ao vivo, ou eventos ao vivo.

18. Produto de programa de computador, caracterizado por ser armazenado em uma mídia de armazenagem legível em computador que quando executada em um processador é disposta para executar as etapas da reivindicação 1.

19. Sistema de busca fonética, caracterizado por ser organizado para executar as etapas da reivindicação 1.

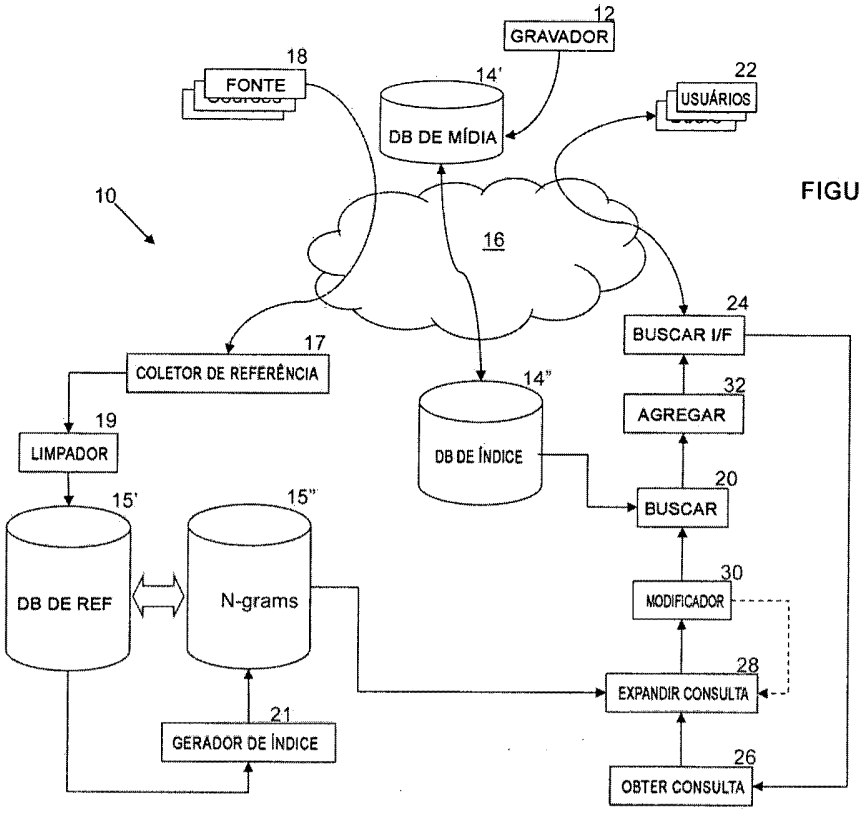

12
GRAVADOR
18
FONTE
14'
DB DE MÍDIA
22
USUÁRIOS
10
16
24
BUSCAR I/F
17
COLETOR DE REFERÊNCIA
14"
32
AGREGAR
19
LIMPADOR
DB DE ÍNDICE
20
BUSCAR
15'
15"
DB DE REF
N-grams
30
MODIFICADOR
28
EXPANDIR CONSULTA
21
GERADOR DE ÍNDICE
26
OBTER CONSULTA

FIGURA 1

RESUMO

SISTEMA E MÉTODO PARA BUSCA FONÉTICA DE DADOS

Um método para foneticamente buscar mídia incluindo uma pluralidade de trilhas de áudio é revelado onde cada trilha de áudio é indexada para fornecer uma representação fonética da trilha de áudio. O método compreende obter uma consulta de busca de texto e buscar a consulta de texto contra um conjunto de documentos de referência para obter um subconjunto de documentos pseudo-relevantes. Os documentos pseudo-relevantes são examinados para um conjunto de expressões de busca caracterizando os documentos pseudo-relevantes. Uma representação fonética correspondendo a pelo menos algumas do conjunto de expressões de busca é fornecida e para cada das representações fonéticas das expressões de busca, as representações fonéticas indexadas para uma ou mais da pluralidade de trilhas de áudio são foneticamente buscadas para fornecer quaisquer indicadores da incidência da expressão de busca em uma ou mais trilhas de áudio. Os indicadores resultantes da busca fonética são combinadas em um conjunto de resultados combinados para cada do conjunto de expressões de busca; e os resultados combinados retornados.