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

(21) PI 0908956-0 A2

(22) Data do Depósito: 19/03/2009

(43) Data da Publicação Nacional: 24/09/2009

(54) Título: COMPONENTE DE SERVIÇO DE CONCEITO DE UM SISTEMA DE BUSCA DE CONTEÚDO E MÉTODO PARA BUSCAR E IDENTIFICAR PONTOS EM UM ITEM DE CONTEÚDO DE MÍDIA

(51) Int. Cl.: G06F 17/30; G06F 17/27; G06F 17/00.

(30) Prioridade Unionista: 19/03/2008 US 12/077,590.

(71) Depositante(es): DELVE NETWORKS, INC.

(72) Inventor(es): JONATHAN THOMPSON; VIJAY CHEMBURKAR; DAVID BARGERON.

(86) Pedido PCT: PCT US2009001782 de 19/03/2009

(87) Publicação PCT: WO 2009/117155 de 24/09/2009

(85) Data da Fase Nacional: 20/09/2010

(57) Resumo: "COMPONENTE DE SERVIÇO DE CONCEITO DE UM SISTEMA DE BUSCA DE CONTEÚDO, MÉTODO PARA BUSCAR E IDENTIFICAR PONTOS EM UM ITEM DE CONTEÚDO DE MÍDIA E MEIO DE ARMAZENAMENTO LEGÍVEL POR COMPUTADOR TENDO UM CONJUNTO DE INSTRUÇÕES PARA REALIZAR O REFERIDO MÉTODO". A presente invenção refere-se a várias modalidades que incluem componentes de serviço de conceito dos sistemas de serviço de busca de conteúdo que empregam antologias e vocabulários preparados para catego- rias particulares de conteúdo em tempos particulares a fim de pontuar transcrições preparadas a partir de itens de conteúdo para permitir um componente de serviço de busca de um sistema de serviço de busca de conteúdo a atribuir estimativas de afinidade de partes de um item de conteúdo com o critério de busca, a fim de apresentar resultados de busca para os clientes do sistema de serviço de busca de conteúdo. O componente de serviço de conceito processa uma solicitação de busca para gerar listas de termos relacionados, e em seguida emprega as listas de termos relacionados para processar transcrições a fim de pontuar transcrições baseado na informação contida nas antologias.

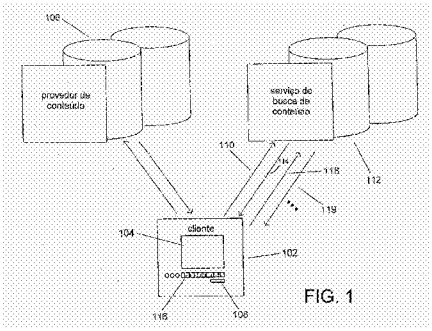


FIG. 1

Relatório Descritivo da Patente de Invenção para **"MÉTODO E SUBSISTEMA PARA BUSCAR CONTEÚDO DE MÍDIA DENTRO DE UM SISTEMA DE SERVIÇO DE BUSCA DE CONTEÚDO".**

REFERÊNCIA CRUZADA PARA PEDIDOS RELACIONADOS

Este pedido é uma continuação em parte do Pedido No 11/903.279, depositado em 21 de setembro de 2007.

Campo da Técnica

A presente invenção refere-se a buscar conteúdo de mídia, incluindo arquivos de vídeo com faixas de áudio, faixas de áudio, e outros tipos de conteúdo de mídia que incluem dados que podem ser total ou parcialmente transcritos para produzir uma transcrição de texto, para identificar partes do conteúdo de mídia relacionado aos termos e frases da consulta de busca, e, em particular, a um componente de serviço de conceito de um sistema de serviço de busca de conteúdo que emprega ontologia e transcrição de texto para pontuar a transcrição para uso por um componente de serviço de busca do serviço de busca de conteúdo para apresentar resultados de busca para um cliente do sistema de serviço de busca de conteúdo.

Fundamentos da Invenção

Nos primórdios da computação, a informação geralmente era codificada como sequências formatadas de caracteres alfanuméricos ou como sequências ordenadas não formatadas de unidades de armazenamento de informação, tipicamente bytes. Conforme o hardware, sistemas operacionais, e aplicações de computador têm evoluído conjuntamente, muitos tipos diferentes de codificações de informação agora são rotineiramente codificados, armazenados, trocados e apresentados eletronicamente para acesso pelos usuários, incluindo arquivos de texto, arquivos específicos de aplicações com formatação especial, gravações de áudio, gravações de vídeo, e apresentações multimídia. Enquanto, nos primórdios da computação, dados eram apresentados primariamente como sequências de caracteres exibidas em terminais monocromáticos de 24 linhas, os muitos tipos diferentes de informação correntemente codificada eletronicamente e distribuída através

de sistemas de computador são apresentados para exibição para usuários humanos através de uma variedade de programas de aplicação diferentes, incluindo editores de texto e imagem, tocadores de vídeo, tocadores de áudio, e navegadores de web.

Uma classe importante de informação compreende informação codificada como uma sequência ordenada de unidades de informação que são apresentadas sequencialmente para exibição ou apresentação para um ser humano. Um vídeo codificado em MPEG é um exemplo de codificação de informação ordenada sequencialmente. Codificação MPEG emprega uma quantidade de camadas bastante complexa de diferentes tipos de codificação e métodos de codificação para codificar compactamente um fluxo de vídeo e/ou fluxo de áudio. Em geral, quadros de vídeo são reconstruídos a partir de um arquivo de vídeo codificado em MPEG quadro a quadro, em sequência. Apresentação de um arquivo de vídeo codificado em MPEG fornece um fluxo de quadros de vídeo e um fluxo de áudio. Aplicações e dispositivos de apresentação geralmente permitem que um usuário inicie ou retome a apresentação do arquivo de vídeo, pare a apresentação do arquivo de vídeo, e salte para frente ou para trás para selecionar posições dentro de um fluxo de vídeo.

Em muitos casos um usuário pode apenas estar interessado em certa parte de uma apresentação de vídeo. Por exemplo, um usuário particular pode estar interessado apenas em um boletim meteorológico incluído em um noticiário de televisão local que inclui recapitulação dos eventos correntes locais e nacionais, recapitulação dos eventos esportivos, e apresentações de histórias de interesse humano adicionalmente ao boletim meteorológico. Em muitos casos, as apresentações de vídeo podem não ser indexadas através de seções, a fim de facilitar o acesso direto a partes da apresentação de vídeo de interesse para um usuário, ou podem ser indexadas em uma granularidade muito grossa de tópicos que requer que um usuário empregue uma estratégia de tentativa e erro de começar, parar, avançar e reverter o fluxo de vídeo através de técnicas relativamente cruas a fim de localizar partes de interesse. Adicionalmente a vídeo codificado, existem muitos

outros tipos de informação ordenada sequencialmente que são apresentados sequencialmente para percepção humana, que incluem gravações de áudio puras, vários tipos de apresentações multimídia, imagens de páginas dentro de livros e documentos de texto, e outras codificações de informação semelhantes. Em muitos casos, buscar partes de informação codificada de interesse para usuários humanos é atualmente limitado as operações de parar/iniciar/avançar e reverter descritas acima familiares aos usuários de aplicações de apresentação de vídeo e muitos dispositivos de apresentação de sinal de vídeo.

Projetistas e fabricantes de computadores e outros dispositivos eletrônicos que apresentam codificações de informação ordenada sequencialmente para exibição para usuários humanos, projetistas, implementadores, vendedores e usuários de aplicações de apresentação de informação, que inclui tocadores de mídia, navegadores de web, e programas de controle, e muitos outros envolvidos na gravação, disseminação e apresentação de informação têm reconhecido a necessidade de ferramentas de busca mais efetivas para permitir que os usuários identifiquem e tenham acesso eficientemente a partes de uma informação codificada de interesse daqueles para quem a informação é apresentada. Em resposta a estas necessidades, tem sido desenvolvido um sistema de serviço de busca de conteúdo. Várias modalidades da presente invenção compreendem componentes e subsistemas deste sistema de serviço de busca de conteúdo. O sistema de serviço de busca de conteúdo recebe e/ou localiza e recupera vários itens de conteúdo disponíveis eletronicamente para clientes do sistema de serviço de busca de conteúdo e prepara representações internas dos itens de conteúdo, ou partes dos itens de conteúdo, para permitir que o serviço de busca de conteúdo apresente graficamente os resultados de busca gerados pelo sistema de serviço de busca de conteúdo em resposta as solicitações de busca feitas pelos clientes do sistema de serviço de busca de conteúdo. Projetistas, desenvolvedores e fabricantes de sistemas de serviço de busca de conteúdo, bem como fornecedores de serviço de busca de conteúdo e usuários de sistemas de serviço de busca de conteúdo de serviços fornecidos através de

serviços de busca de conteúdo, têm todos reconhecido a necessidade por componentes de serviço de busca de conteúdo eficientes e precisos para facilitar repostas rápidas e precisas às solicitações de busca direcionadas a itens de conteúdo recebidos dos clientes de serviços de busca de conteúdo que empregam sistemas de serviço de busca de conteúdo.

Sumário da Invenção

Várias modalidades da presente invenção incluem componentes de serviço de conceito dos sistemas de serviço de busca de conteúdo que empregam ontologias e vocabulários preparados para categorias particulares de conteúdo em tempos particulares a fim de pontuar transcrições preparadas a partir de itens de conteúdo para habilitar um componente de serviço de busca de um sistema de serviço de busca de conteúdo a atribuir estimativas de relacionamento de partes de um item de conteúdo com o critério de busca a fim de apresentar resultados de busca para os clientes do sistema de serviço de busca de conteúdo. O componente de serviço de conceito processa uma solicitação de busca para gerar listas de termos relacionados, e em seguida emprega as listas de termos relacionados para processar transcrições a fim de pontuar transcrições baseado na informação contida nas ontologias.

Breve Descrição das Figuras

A Figura 1 ilustra o fornecimento de serviços de busca para um cliente através de um sistema de serviço de busca de conteúdo.

A Figura 2 ilustra uma interface de aplicação de apresentação de conteúdo.

A Figura 3 ilustra uma exibição de mapa de calor de resultados de busca que é adicionado à interface de aplicação de apresentação de conteúdo mostrada na Figura 2 de acordo com as modalidades da presente invenção.

A figura 4 fornece uma representação de diagrama de bloco de um sistema de serviço de busca de conteúdo que representa uma modalidade da presente invenção.

A Figura 5 ilustra uma ontologia de acordo com uma modalidade

da presente invenção.

A Figura 6 mostra uma parte de um vocabulário ilustrativo para a categoria "animais" de acordo com uma modalidade da presente invenção.

A Figura 7 ilustra uma representação *m* x *m* de uma ontologia, de acordo com uma modalidade da presente invenção.

A Figura 8 mostra uma representação de lista de uma ontologia, de acordo com uma modalidade da presente invenção.

A Figura 9 mostra uma parte de uma representação de ontologia usada em uma modalidade da presente invenção.

A Figura 10 ilustra uma implementação de uma transcrição, de acordo com uma modalidade da presente invenção.

A Figura 11 ilustra uma representação alternativa de uma transcrição, de acordo com uma modalidade da presente invenção.

A Figura 12 ilustra operação do componente de serviço de conceito (408 na Figura 4) de um sistema de serviço de busca de conteúdo que representa uma modalidade da presente invenção.

A Figura 13 ilustra uma modalidade de uma transcrição pontuada, de acordo com uma modalidade da presente invenção.

A Figura 14 fornece um diagrama de fluxo de controle para o componente de serviço de conceito de um sistema CSS, de acordo com uma modalidade da presente invenção.

A Figura 15 é um diagrama de fluxo de controle para a rotina "processar frase de busca" chamada na etapa 1408 da figura 14, de acordo com uma modalidade da presente invenção.

A Figura 16 fornece um diagrama de fluxo de controle para a rotina "processar transcrição" chamada na etapa 1412 da Figura 14, de acordo com uma modalidade da presente invenção.

Descrição Detalhada da Invenção

Modalidades da presente invenção são empregadas dentro de sistemas de serviço de busca de conteúdo que são usados para fornecer serviço de busca de conteúdo para clientes. A Figura 1 ilustra o fornecimento de serviços de busca para um cliente através de um sistema de serviço de

busca de conteúdo. O cliente 102 é geralmente um computador pessoal ou estação de trabalho empregada por um usuário para visualizar conteúdo 104, fornecido através de um sistema provedor de conteúdo 106, através de uma aplicação de apresentação de conteúdo, tal como um programa de apresentação de vídeo chamado por um navegador web. A fim de facilitar a visualização eficiente do conteúdo, o usuário entra com uma consulta de busca, que compreende um termo de busca ou uma frase de busca, em um recurso de entrada de texto 108 exibido no computador cliente e chama uma busca do conteúdo que é apresentado para partes relacionadas do conteúdo. Em certas modalidades da presente invenção, consultas de busca podem alternativamente ser predefinidas através ou pelos usuários para facilitar a busca do usuário. Os termos ou frases de busca são transmitidos 110 para um sistema de serviço de busca de conteúdo 112 e processados pelo sistema de serviço de busca de conteúdo a fim de retornar 114 uma apresentação gráfica das partes particulares relacionadas do conteúdo 116 para o termo de busca ou frase de busca. O usuário do computador cliente pode executar buscas adicionais para termos ou frases adicionais através de transações de busca adicionais 118 e 119.

Em geral, o conteúdo apresentado ao usuário é um tipo de conteúdo que, diferente de um arquivo de texto, não pode ser facilmente pesquisado com o uso de ferramentas de busca disponíveis usualmente, tais como mecanismos de busca fornecidos dentro de editores de texto, a fim de encontrar partes do conteúdo de interesse particular para o usuário. Na discussão a seguir, é assumido que o conteúdo é um arquivo de vídeo com uma faixa de áudio associada, tal como um noticiário ou transmissão esportiva fornecidos através de um serviço de notícias ou serviço de esportes através da Internet para os usuários que fazem acesso. Entretanto, os sistemas de serviço de busca de conteúdo, nos quais podem ser implementadas e empregadas modalidades da invenção corrente, podem fornecer serviços de busca para uma variedade de tipos de conteúdo diferentes, desde apresentações multimídia até vários tipos de seleções de imagens, gráficos, e musicais.

De modo geral, o provedor de conteúdo 106 fornece itens de conteúdo para o sistema de serviço de busca de conteúdo 112 para pré-processamento, para facilitar respostas rápidas para as solicitações de busca de cliente subsequentes direcionadas para os itens de conteúdo. Entretanto, em implementações alternativas, o sistema de serviço de busca de conteúdo pode receber concorrentemente o item de conteúdo e frase de busca ou item de busca a partir de um cliente, processar o item de conteúdo para preparar para a busca de conteúdo, executar a solicitação de busca e apresentar os resultados para o usuário em tempo real. De modo geral, os sistemas clientes são distintos tanto dos sistemas fornecedores de conteúdo como dos sistemas de serviço de busca de conteúdo, embora seja possível que o provedor de conteúdo, cliente e serviço de busca de conteúdo possam todos ser executados concorrentemente ou simultaneamente dentro de um único sistema computador ou sistema de computador distribuído.

A Figura 2 ilustra uma interface de aplicação de apresentação de conteúdo. Vídeo é exibido dentro da tela de vídeo 210 fornecido por uma interface de visualização de clipe de vídeo baseada em página da web ou interface gráfica de usuário de dispositivo portátil ("GUI") 212. A interface do dispositivo ou página da web fornece uma janela de entrada de texto 214 que permite que um usuário entre texto para servir como critério de busca para encontrar os vídeos desejados para visualizar, exibir os resultados de cada busca em uma janela de resultados 216 que pode ser rolada através de botões de rolagem para cima e de rolagem para baixo e a partir da qual pode ser selecionado vídeo para exibição. Adicionalmente, é exibida um imagem de indicação de progresso 222, para um usuário, uma indicação da posição atual dentro de um clipe de vídeo que está sendo exibido durante a apresentação do clipe de vídeo, com a duração inteira do clipe de vídeo representada por uma barra horizontal 224 e a posição corrente dentro do clipe de vídeo indicada pela posição de um indicador de posição 226 com respeito à barra horizontal. Na Figura 2, o indicador de posição 226 indica que o quadro de vídeo exibido atualmente ocorre a uma posição de 25% do curso do clipe de vídeo. A interface de usuário fornece um botão de iniciar/parar

228 para iniciar e parar a exibição de um clipe de vídeo, bem como um botão de retrocesso 230 e botão de avanço 232 que permitem que o usuário procure posições diferentes dentro do clipe de vídeo sem assistir os quadros que passam.

A Figura 3 ilustra uma exibição de um mapa de calor de resultados de busca que é adicionado à interface de aplicação de apresentação de conteúdo mostrada na Figura 2 de acordo com modalidades da presente invenção. A exibição de mapa de calor de resultados de busca pode ser fornecida através de uma aplicação do lado cliente da apresentação de resultados de busca baixada a partir de um sistema de serviço de busca de conteúdo. A Figura 3 mostra os botões de navegação e exibição de progresso de uma interface de visualização de clipe de vídeo mostrada na Figura 2, juntamente com recursos adicionais de exibição de resultados de busca. Os botões de retrocesso 130, avanço 132 e iniciar/parar 128 têm funções idênticas na interface visual às funções descritas para estes recursos de interface da interface de visualização de clipe de vídeo na Figura 2. A exibição do progresso 124 e 126 também tem uma função idêntica àquela da interface de visualização de clipe de vídeo mostrada na Figura 2, com exceção de que, em lugar de mostrar uma barra horizontal de cor sólida simples para representar a extensão do clipe de vídeo, como na Figura 2, a representação semelhante a um mapa de calor de uma função relacionada é sobreposta dentro da barra horizontal 124 do progresso da exibição. Nesta representação semelhante a mapa de calor, tonalidade mais escura representa métrica ou pontuação relacionada de maior magnitude. A interface visual também inclui dois recursos de especificação de parâmetro 302 e 304 que permitem que um usuário especifique, através de deslizar os botões indicadores 306 e 308, respectivamente ao longo das colunas 310 e 312, o grau de limitação e suavização a empregar quando computa as métricas ou pontuações relacionadas para posições dentro de uma informação codificada com respeito a um critério de busca 314 especificado pelo usuário dentro de uma janela de entrada de critério de busca 316. No exemplo mostrado na Figura 3, a barra horizontal 14 do componente de exibição de progresso representa a exten-

são de um clipe de vídeo, e uma pessoa pode determinar facilmente, através de inspeção visual do mapa de calor sobreposto dentro da barra horizontal 124, que o conteúdo relacionado ao critério de busca especificado correntemente pode ser achado nas posições 320, 322 e 324 com maiores probabilidades. Uma interface visual mais simples pode incluir apenas uma representação semelhante a mapa de calor de uma função relacionada, e pode contar com recursos de seleção de uma GUI existente para entrar com critérios de busca. Interfaces visuais mais complexas podem incluir recursos de seleção adicionais para permitir parâmetros adicionais que controlem a exibição da interface visual e computação da função de relacionabilidade a ser especificada por um usuário, incluindo domínio do argumento, por exemplo. Naturalmente, como com todas as interfaces visuais, existem muitas formas diferentes, e tipos de seleção e recursos de entrada, que podem ser usados para fornecer entrada de usuário de parâmetros, critérios de busca e outros dados de entrada. Adicionalmente, uma interface visual pode suportar múltiplos métodos para dar entrada a quaisquer dados de entrada particulares. Por exemplo, na interface visual mostrada na Figura 3, um usuário pode estar apto a selecionar uma posição na qual iniciar ou retomar a apresentação de informação codificada através do uso dos botões de retroceder e avançar, através de mover o indicador de posição, ou através de dar entrada a um clique de mouse depois de mover um ponteiro para a posição como representada por uma localização dentro da barra horizontal do componente de progresso de exibição.

A Figura 4 fornece uma representação de diagrama de bloco de um sistema de serviço de busca de conteúdo que representa uma modalidade da presente invenção. O sistema de serviço de busca ("sistema CSS") 402 inclui um componente de serviço de busca 404 que recebe solicitações de busca a partir de clientes e responde com resultados que são apresentados através de aplicações de apresentação de resultados de busca que rodam em computadores clientes. Em uma modalidade da presente invenção, as solicitações de busca e respostas às solicitações de busca são recebidas 406 e transmitidas 407 pela Internet de acordo com o protocolo de serviços

web nas mensagens de linguagem de marcação extensível XML. Uma solicitação de busca inclui um identificador de conteúdo ("ID de Conteúdo") em uma consulta de busca, que compreende um termo de busca ou frase de busca. Ambos estes itens são passados, pelo componente de serviço de busca 404, para um componente de serviço de conceito ("Componente CS") 408 para processamento. O componente CS 408 retorna uma transcrição pontuada 404, ou partes pontuadas de uma transcrição, para o componente de serviço de busca 404, que usa a transcrição pontuada, e, opcionalmente, um vocabulário retornado pelo componente de serviço de conceito, para produzir informação do resultado de busca que é apresentado a um usuário no computador cliente. Uma transcrição pontuada, ou transcrição pontuada parcialmente, é, nas modalidades descritas da presente invenção, uma lista de tuplas ordenada por tempo, em que cada tupla contém uma indicação de um termo ou uma frase, o tempo decorrido, durante a apresentação de um item de conteúdo de mídia, no qual o termo ou frase ocorre, e uma pontuação que indica o grau de relacionabilidade do termo ou frase a consulta de busca recebida pelo componente CS. O componente CS faz acesso a um armazenamento de ontologia 414, um componente de dados de conteúdo extraído 416, e um serviço de dados de mídia 418 a fim de obter uma ontologia, uma transcrição, e outra informação necessária pelo componente CS para pontuar uma transcrição para retornar para o componente de serviço de busca 404. Uma ontologia é nas modalidades descritas da presente invenção, um gráfico totalmente interconectado de termos e frases. Cada nó do gráfico representa um termo ou uma frase, e cada borda do gráfico representa o relacionamento de co-ocorrência de termos e frases representados pelos nós conectados pela borda dentro da informação coletada que é analisada para produzir a ontologia. A cada borda é atribuído um peso que reflete a força do relacionamento de co-ocorrência representado pela borda, e os pesos são derivados da informação coletada que é analisada para produzir a ontologia. O armazenamento de ontologia 414 inclui uma quantidade de ontologias tal como a ontologia 422, que descreve relacionamentos de co-ocorrência entre palavras para várias categorias do argumento. As ontologi-

as também têm a data registrada, ou data/hora registrada, uma vez que as ontologias mudam, ao longo do tempo, para algum argumento particular, e uma ontologia com um registro de data/hora que indica uma data dentro de um deslocamento razoável em tempo, dos dados de um item de conteúdo a ser pesquisado é mais útil para preparar os resultados de busca. O componente de dados de conteúdo extraído 416 armazena uma ou mais transcrições 426 para cada item de conteúdo que tenha sido pré-processado pelo sistema CSS. O serviço de dados de mídia 418 armazena informação relacionada a cada item de conteúdo pré-processado, que inclui a categoria do argumento ao qual o item de conteúdo pertence e a data ou data e hora de criação ou recepção do conteúdo.

O serviço CSS adicionalmente inclui um componente de serviço de conteúdo 430 que recebe itens de conteúdo através de fornecedores de conteúdo remotos, e fornece os itens de conteúdo para um componente processador de conteúdo 432 que prepara a armazena uma ou mais transcrições 426 para cada item de conteúdo processado no componente de dados de conteúdo extraído 416. O processador de conteúdo 432 acessa um modelo de linguagem, tal como modelo de linguagem 434, armazenado em um armazenamento de modelo de linguagem 436, a fim de processar um dado item de conteúdo. O componente processador de conteúdo 432 também deposita informação adicional sobre itens de conteúdo no componente de serviço de dados de mídia 418. Nas modalidades descritas da presente invenção, transcrições são transcrições baseadas em texto de faixas de áudio e arquivos de áudio, executadas através de sub-componentes de reconhecimento de voz automáticos do componente processador de conteúdo. Em modalidades alternativas da presente invenção, transcrições de texto podem ser preparadas a partir de outros tipos de conteúdo de mídia, que inclui transcrições descritivas imagens imóveis ou móveis preparadas através de sub-componentes de percepção visual do computador do componente processador de conteúdo.

Um componente classificador e agregador de informação 440 busca continuamente, ou em intervalos, através de informação disponível na

Internet e outras fontes de informação por documentos, arquivos de texto, e outros itens de informação relacionados a várias categorias as quais os itens de conteúdo podem ser vinculados. O componente classificador e agregador de informação 440 classifica aqueles itens de informação que se acredita sejam úteis para o sistema CSS por categoria, e armazena os itens de informação, para cada categoria a para faixas particulares de datas e horas, dentro de um componente de armazenamento de informação categorizada 442. Estes itens de informação são processados pelo componente classificador e agregador de informação para remover informação desnecessária, normalizar linguisticamente termos e frases, e computar vários parâmetros e valores associados com os itens de informação que são usados tanto pelo componente classificador e agregador de informação para classificar os itens como pelo componente construtor de modelo de linguagem 444 e componente construtor de ontologia 446 que usa itens de informação armazenados no componente de armazenamento de informação categorizada 442 para construir modelos de linguagem e ontologias respectivamente.

A Figura 5 ilustra uma ontologia de acordo com uma modalidade da presente invenção. A Figura 5 é uma ontologia simplificada que contém apenas uns poucos termos. Ontologias reais preparadas para categorias de informação úteis podem conter muitas centenas, milhares, ou milhões de termos e frases. Na Figura 5, cada um dos seis termos é representado por nós ovais, tal como o nó oval 502 que representa o termo "cobra". Cada par de termos possíveis, tal como o par de termos "cobra" 502 e "pele" 504, são interconectados através de dois arcos, tais como os arcos 506 e 508 que interconectam os termos 502 e 504. Os dois arcos formam um par bidirecional, um arco do par direcionado de um primeiro termo ou frase (termo fonte ou frase fonte para o arco) para um segundo termo ou frase (termo alvo ou frase alvo para o arco), e o segundo arco do par direcionado do segundo termo ou frase para o primeiro termo ou frase. Cada arco é rotulado com um valor numérico na faixa [0,0; 1,0]. O valor numérico é uma métrica de co-ocorrência normalizada que indica uma frequência na qual o termo ou frase alvo do arco co-ocorre com o termo ou frase fonte do arco. Deste modo, na

Figura 5, o arco 506 indica que o termo "cobra" co-ocorre a uma frequência relativamente baixa com o termo "pele", enquanto o termo "pele" co-ocorre a uma frequência um pouco maior com o termo "cobra". O fato de que as métricas de co-ocorrência para os dois arcos em um par bidirecional de arcos que interconectam dois termos ou frase não são iguais reflete distribuições diferentes de termos ou frases e quantidades diferentes de ocorrências de termos ou frases nos muitos itens de informação a partir dos quais as ontologias são preparadas, bem como com diferentes normalizações para os dois termos ou frases. Com referência de novo a Figura 4, as ontologias, tal como a ontologia simples mostrada na Figura 5, são preparadas pelo componente construtor de ontologias 446 do sistema CSS através de análise de uma grande quantidade de itens de informação relacionados a uma categoria particular e coletados por um intervalo de tempo particular. Deste modo, cada ontologia, tal como a ontologia ilustrada na Figura 5, é associada com uma categoria particular e é marcada com uma data e ou data/hora que correspondem à data ou data e hora, respectivamente, quando as entidades de informação usadas pelo componente construtor de ontologia para construir a ontologia foram coletados pelo componente classificador e agregador de informação 4450 do sistema CSS 402.

Cada ontologia é física ou conceitualmente associada com um vocabulário. O vocabulário também é preparado a partir de itens de informação coletados pelo componente classificador e agregador de informação (440 na Figura 4) do sistema CSS. Em certas modalidades, o vocabulário para uma categoria de informação é preparado pelo componente construtor de modelo de linguagem (444 na Figura 4) do sistema CSS e armazenado no armazenamento de modelo de linguagem (436 na Figura 4). Em outras modalidades da presente invenção, o vocabulário pode ser construído pelo componente construtor de ontologia (446 na Figura 4) e armazenado no armazenamento de ontologia (414 na Figura 4), e ainda em modalidades alternativas, o vocabulário pode ser construído por também um componente adicional do CSS.

Um vocabulário compreende uma lista de substantivos, ou fra-

ses de substantivos, em uma modalidade da presente invenção, isto ocorre usualmente em itens de informação relacionadas a uma categoria de informação particular. Por exemplo, uma categoria de esportes para itens de conteúdo pode se esperar que inclua substantivos tais como "bastão", "base", "arremessador", "lançador", "trave", "futebol"," dardo", "patinação", e outros substantivos e frases de substantivos. Devido ao fato de ser ineficiente manipular programaticamente sequências de símbolos, tais como sequências de caracteres, quando se implementa componentes do sistema CSS, cada termo ou frase em um vocabulário é representado por um valor inteiro. A Figura 6 mostra uma parte de um vocabulário ilustrativo para a categoria "animais", de acordo com uma modalidade da presente invenção. Como pode ser visto na Figura 6, a representação da sequência de caracteres do nome de cada animal, tal como a sequência de caracteres "aardvark" 602, é associada com um pequeno valor inteiro, tal como valor "96" 604 na tabela 606 que constitui um vocabulário para a categoria de informação "animais". Usando esta tabela, a sequência de caracteres "jacaré" 608 é facilmente traduzida para o inteiro "462" 610 através de uma operação de pesquisa na tabela. Como qualquer dado processado computacionalmente e armazenado eletronicamente, o vocabulário pode ser adicionalmente associado com índices ou outra informação adicional para permitir termos e frases para serem localizados rapidamente na tabela e acessados.

Ao mesmo tempo em que é conveniente representar uma ontologia como um gráfico que inclui nós de termos e frases interconectados por arcos, como mostrado na Figura 5, uma ontologia pode ser manipulada mais facilmente, computacionalmente, quando representada como uma matriz *m x m*, onde *m* é a quantidade de termos e frases de um vocabulário particular. A Figura 7 ilustra uma representação *m x m* de uma ontologia, de acordo com uma modalidade da presente invenção. A matriz *m x m* 702 compreende m^2 células, em que cada célula, tal como a célula 704, contém uma ou mais métricas de co-ocorrência que rotulam um arco, tal como o arco 508 na Figura 5, que emana de um primeiro nó de ontologia, tal como o nó 502 na Figura 5, e direcionado para um segundo nó de ontologia, tal como nó 504 na Figu-

ra 5. O índice da linha da célula indica o valor inteiro que corresponde ao primeiro nó, a partir do qual o arco emana, e o índice da coluna da célula indica o segundo nó, para o qual o arco é direcionado. A célula 704 tem índices de matriz (5, *m*-1), que indicam que as métricas de co-ocorrência incluídas na célula, tal como métrica "0,20" 706 na Figura 7, rotulam um arco de uma palavra ou frase de vocabulário especificada pelo inteiro "5" até o termo ou frase do vocabulário especificado pelo inteiro *m*-1.

A representação *m x m* de uma ontologia, mostrada na Figura 7, é uma abstração útil, mas também é de forma geral ineficiente computacionalmente. Uma razão pela qual esta representação é ineficiente é que, para ontologias práticas, as métricas de co-ocorrência abaixo de um valor limite são consideradas sem significado, e a todas é atribuído um valor mínimo, tal como o valor "0,0". Portanto, a matriz *m x m*, mostrada na Figura 7, é geralmente bastante dispersa. Por este motivo, a para facilitar acesso rápido a métricas de co-ocorrência particulares para palavras e frases particulares do vocabulário, a ontologia é usualmente representada como uma lista. A Figura 8 mostra uma representação de lista da uma ontologia, de acordo com uma modalidade da presente invenção. Na Figura 8, cada elemento da lista 802, tal como o elemento 804, é representado como uma linha que contém três células. A primeira célula 806 da linha 804 é a representação numérica do alvo de um arco na representação gráfica de uma ontologia, a segunda célula 808 é a fonte de um arco, na representação gráfica de uma ontologia, e a terceira célula 810 contém a métrica de co-ocorrência pela qual o arco é rotulado. Somente entradas com métricas diferentes de zero são incluídas na lista 802, o que resolve o problema de dispersão associado com representações *m x m* de uma ontologia. Cada entrada na lista representa um único arco de uma ontologia. As entradas são ordenadas, na Figura 8, em ordem ascendente com respeito ao valor armazenado na primeira célula de cada entrada, conforme visto prontamente pelos valores nas primeiras células das entradas na Figura 8. Esta organização facilita o acesso a aquelas entradas associadas com um termo ou frase particular ao qual um arco é direcionado na representação gráfica da ontologia. Em certas modalidades,

as entradas podem ser ordenadas adicionalmente com respeito ao valor armazenado na segunda célula de cada entrada, e ainda em modalidades adicionais, a representação da lista de uma ontologia pode ser acompanhada por uma ou mais tabelas de referência, ou índices, para facilitar o acesso rápido a entradas particulares da ontologia.

Na prática, mesmo a representação da lista de uma ontologia, mostrada na Figura 8, pode ser de alguma forma uma abstração. Em uma modalidade da presente invenção, a ontologia inclui os dados da linha empregados para computar a métrica da co-ocorrência, para cada entrada, em vez de a métrica de co-ocorrência computada. A Figura 9 mostra uma parte de uma representação de ontologia usada em uma modalidade da presente invenção. A ontologia é representada como uma lista 902, similar a representação da lista ilustrada na Figura 8. Entretanto, em vez de incluir uma métrica de co-ocorrência computada única, como na Figura 8, cada entrada na lista da Figura 9 inclui, em uma modalidade da presente invenção, três valores numéricos 904 a 906 que codificam a quantidade de ocorrências da palavra ou frase representada pelo valor armazenado no primeiro elemento 908 da entrada, dentro do mesmo item de informação, ou dentro de uma subunidade ou subseção do item de informação, uma vez que a palavra ou frase representada pelo valor armazenado na segunda célula 910 da entrada em uma grande quantidade de itens de informação coletados e processados que corresponde a categoria da informação para qual a ontologia é preparada.

Na presente discussão, itens de conteúdo são arquivos de vídeo que incluem faixas de áudio. Em uma modalidade da presente invenção, a busca é executada pelo sistema CSS exclusivamente na faixa de áudio de um arquivo de vídeo, usando termos e frases entrados por um usuário para encontrar aqueles termos ou frases, ou termos e frases relacionados, que ocorrem em pontos no tempo na faixa de áudio. Deste modo, partes da faixa de áudio podem ser identificadas como sendo relacionadas aos termos de busca e de interesse particular para um usuário. Aquelas partes da faixa de áudio podem, por sua vez, ser relacionadas às imagens de vídeo que são

exibidas no intervalo de tempo no qual as partes da faixa de áudio são apresentadas, quando o arquivo de vídeo é apresentado para o usuário através de uma aplicação de apresentação de arquivo de vídeo. Nestas modalidades, uma transcrição (426 na Figura 4) é essencialmente uma lista de ocorrências de termos ou frases associados com um tempo, ou intervalo de tempo, em que os termos ou frases dos termos ocorrem na faixa de áudio durante a apresentação da faixa de áudio para um usuário. A Figura 10 ilustra uma implementação de uma transcrição, de acordo com uma modalidade da presente invenção. Na Figura 10, cada célula em uma matriz unidimensional 1002, tal como uma célula 1004, ou é branca, que indica que nenhuma palavra ou frase foi reconhecida durante aquele intervalo de tempo, ou contém uma representação numérica de uma palavra ou frase selecionada a partir de um vocabulário associado com a categoria do item de conteúdo a partir do qual a transcrição é preparada. Nesta modalidade de uma transcrição, cada célula representa um pequeno intervalo de tempo fixo, de modo que a matriz unidimensional 1002 representa uma linha de tempo para apresentar a faixa de áudio de um arquivo de vídeo. A Figura 11 ilustra uma representação alternativa de uma transcrição, de acordo com uma modalidade da presente invenção. Na Figura 11, a transcrição é representada como uma lista, ou matriz bidimensional, cada entrada, ou linha, da qual contém um valor numérico que indica uma palavra ou frase a partir de um vocabulário, tal como um valor numérico 1102, e um tempo associado no qual a palavra ou frase ocorrem na faixa de áudio, tal como tempo 1104, ambos dentro da entrada 1106, Muitas representações alternativas de transcrições são possíveis.

A Figura 12 ilustra a operação do componente CS (408 na Figura 4) de um CSS que representa uma modalidade da presente invenção. O componente CS recebe um ID de conteúdo 1202 em uma consulta de busca 1203 a partir de um componente de serviço de busca (404 na Figura 4) do CSS que representa uma modalidade da presente invenção. O ID de conteúdo é geralmente um identificador numérico, ou sequência alfanumérica, que identifica unicamente um item de conteúdo particular. O componente CS usa

o ID de conteúdo 1202 para acessar o componente de serviço de dados de mídia (418 na Figura 4) para obter um ID de categoria 1204 para o item de conteúdo e uma data/hora 1206 para o item de conteúdo. O componente CS adicionalmente acessa o armazenamento de ontologia (414 na Figura 4) e, em certas modalidades, o armazenamento de modelo de linguagem (436 na Figura 4) a fim de obter uma ontologia 1208 e um vocabulário 1210 apropriados para o item de conteúdo. Usando a ontologia e vocabulário 1208 e 1210, e usando várias regras e rotinas de processamento de linguagem, o componente CS então processa a consulta de busca recebida 1203 para gerar uma ou mais listas de termos ou frases 1212 e 1214. Primeiramente, a consulta de busca tem os erros de grafia corrigidos e é parcialmente normalizada para produzir um termo ou frase de busca modificado 1216. A consulta de busca modificada 1216 é em seguida processada para extrair aquelas palavras que ocorrem no vocabulário para a categoria a qual o item de conteúdo identificado pelo ID de conteúdo 1202 pertence. A categoria é identificada pelo ID de categoria 1204 obtido a partir do componente de serviço de dados de mídia. Cada lista 1212 e 1214 compreende um termo ou frase de busca e frases e termos de busca relacionados adicionais, como obtido a partir da ontologia 1208. Cada termo ou frase na lista é associado com um valor de métrica de co-ocorrência extraído a partir da ontologia. No exemplo mostrado na Figura 12, os termos "gasolina", "carro", "cobra", e "pele" são encontrados, na ontologia a ser relacionada ao termo de busca "óleo", e, portanto são incluídos na lista 1212 para o termo de busca "óleo". De maneira similar, a lista 1214 contém o termo de busca "carro" e os termos adicionais relacionados "gasolina" e "óleo". Os termos e frases relacionados são obtidos, a partir da ontologia, a partir daquelas entradas nas quais um termo ou frase de consulta de busca ocorre como o primeiro valor nas entradas de ontologia (ver Figuras 8 e 9). Uma vez que a lista tenha sido preparada, o componente CS em seguida acessa o componente de dados de conteúdo extraídos (416 na Figura 4) para obter uma transcrição para o item de conteúdo 1218. O componente CS em seguida usa a lista 1212 e 1214 para atribuir métricas de co-ocorrência para aqueles termos e frases da transcrição

1218 que ocorrem no vocabulário da categoria a qual o item de conteúdo pertence, para produzir uma transcrição pontuada 1220. O serviço CS então, em uma modalidade da presente invenção, devolve a transcrição pontuada e o ID de conteúdo, e, opcionalmente, a frase de busca modificada 1216 e uma referência ao vocabulário, para o componente de serviço de busca (404 na Figura 4) do sistema CSS. O componente de serviço de busca em seguida processa adicionalmente a transcrição pontuada para apresentar os resultados da busca para um usuário.

A Figura 13 ilustra uma modalidade de uma transcrição pontuada, de acordo com uma modalidade da presente invenção. A transcrição pontuada é uma lista de tuplas, da qual cada tupla é representada na Figura 13 por uma linha, tal como a linha 1304. Cada tupla, tal como a tupla 1304, inclui a representação numérica de uma palavra ou frase, uma indicação do tempo decorrido no qual a palavra ou frase ocorrem na transcrição de áudio, e uma pontuação computada para a palavra ou frase. Em geral, a pontuação é uma função da métrica ou métricas de co-ocorrência obtidas a partir da ontologia usada para pontuar a transcrição. Em uma modalidade da presente invenção, por exemplo, a pontuação é simplesmente a métrica de co-ocorrência obtida a partir da ontologia, a menos que o termo ou frase para o qual a pontuação é computada ocorra em múltiplas listas, tal como em ambas as listas 1212 e 1214 no exemplo da Figura 12, em cujo caso a pontuação pode ser computada como a média, ou uma média ponderada, das métricas de co-ocorrência associadas com o termo em qualquer das listas na qual o termo ou frase ocorrem.

A Figura 14 fornece um diagrama de controle para o componente CS de um sistema CSS, de acordo com uma modalidade da presente invenção. Na etapa 1402, o componente CS recebe um ID de conteúdo que identifica unicamente um item de conteúdo e um termo ou frase de busca. Na etapa 1404, o componente CS usa o ID de conteúdo para obter um ID de categoria e data/hora para o ID de conteúdo. O ID de conteúdo identifica a categoria da informação a qual o item de conteúdo pertence, e a data/hora identifica a data ou data e hora com a qual o item de conteúdo está associa-

do, para habilitar o componente CS para encontrar uma ontologia e vocabulário apropriados para o item de conteúdo. Na etapa 1406, o componente CS usa a categoria e data/hora obtidos na etapa 1404 para obter uma ontologia e vocabulário apropriados para o item de conteúdo. Na etapa 1408, o componente CS processa o termo de busca ou frase de busca recebido com o uso da ontologia e vocabulário obtidos, através de uma chamada a rotina "processar frase de busca", chamada na etapa 1408, e uma referência ao vocabulário obtido na etapa 1406.

A Figura 15 é um diagrama de fluxo de controle para a rotina "processar frase de busca" chamada na etapa 1408 da Figura 14, de acordo com uma modalidade da presente invenção. Na etapa 1502, a rotina "processar a frase de busca" recebe uma frase de busca. Na etapa 1504, a grafia dos termos na frase de busca é corrigida e as palavras de frase de busca são normalizadas de acordo com as regras de linguagem e rotinas de linguagem. Por exemplo, termos plurais podem ser substituídos por termos singulares, e termos e frases derivados de termos e frases raiz podem ser substituídos por termos e frases raiz. Então, na etapa 1506, quaisquer termos e frases que não podem ser encontrados no vocabulário obtido na etapa 1406 da Figura 14 são removidos, deixando um ou mais termos e frases selecionados a partir do vocabulário associados com a categoria de informação a qual o item de conteúdo pertence. Então, no loop-*condicional* das etapas 1508 a 1513, é criada uma lista de termos relacionados para cada termo e frase dos termos e frases restantes após a etapa 1506. Novamente, como discutido acima, a pontuação da co-ocorrência associada com cada termo e frase em cada lista é geralmente a métrica de co-ocorrência obtida a partir da ontologia obtida na etapa 1406 da Figura 14.

A Figura 16 fornece um diagrama de fluxo de controle para a rotina "processar transcrição", chamada na etapa 1412 da Figura 14, de acordo com uma modalidade da presente invenção. Na etapa 1602 é criada uma nova lista de termo/hora/pontuação, tal como aquela mostrada na Figura 13. Então, no loop-*condicional* das etapas 1604 a 1607, cada termo na transcrição obtida na etapa 1410 da Figura 14 é considerado durante cada iteração

do loop-*condicional*, e uma tupla de termo/hora/pontuação é entrada na nova lista termo/hora/pontuação, criada na etapa 1602, para o termo ou frase considerado correntemente. Como discutido acima, a pontuação entrada para um termo ou frase é geralmente uma função da métrica ou métricas de co-ocorrência obtida a partir da ontologia, ou, quando o termo ou frase ocorre em múltiplas listas preparadas nos loops-*condicionais* das etapas 1508 a 1513 da Figura 15, a pontuação pode ser computada como uma média, média ponderada, ou alguma outra função de múltiplas ocorrências do termo ou frase e armazenada para o termo ou frase. Em certas modalidades da presente invenção, um cálculo adicional opcional pode ser realizado com a lista de tuplas de termo/hora/pontuação produzida pelos loops-*condicionais* das etapas 1604 a 1607. Por exemplo, em uma modalidade da presente invenção, um loop-*condicional* das etapas 1610 a 1613 pode ser executado para considerar novamente cada tupla termo/hora/pontuação em uma lista termo/hora/pontuação recém criada a fim de modificar cada pontuação de acordo com os termos e frases vizinhos a um dado termo ou frase, no tempo, dentro da lista termo/hora/pontuação. Por exemplo, o fato de que um intervalo de tempo particular na transcrição contém ocorrências de frases ou termos de todos, ou uma maioria, das listas, preparadas nas etapas 1508 a 1523 da Figura 15, pode indicar que as pontuações associadas com os termos e frases naquele intervalo devem ser aumentadas, para refletir uma maior probabilidade de que as ocorrências dos termos e frases sejam realmente relacionados à frase de busca. Muitas considerações adicionais podem ser feitas em passagens adicionais pela lista termo/hora/pontuação. Finalmente, na etapa 1614, a lista termo/hora/pontuação, preparada pela rotina "processar transcrição", é retornada como a transcrição pontuada.

Embora a presente invenção tenha sido descrita em termos de modalidades particulares, não se entende que a invenção esteja limitada a estas modalidades. Ficarão evidentes modificações dentro do espírito da invenção para os indivíduos versados na técnica. Por exemplo, o componente CS de um sistema CSS pode ser implementado em qualquer quantidade de linguagens de programação para execução em qualquer quantidade de

sistemas operacionais diferentes sendo executados em diferentes plataformas de hardware dentro de muitos tipos diferentes de sistemas CSS. Implementações do componente CS podem variar de acordo com variações em parâmetros e características familiares de programação, que incluem estruturas de controle, estruturas de dados, organização modular, e outros parâmetros e características familiares. Como discutido acima, muitos tipos diferentes de ontologias e representações de ontologia, e muitos tipos diferentes de transcrições e representações de transcrição podem ser empregados por várias modalidades do componente CS para preparar transcrição pontuada. As métricas de co-ocorrência e outros valores numéricos podem ter diferentes amplitudes e representações, em modalidades alternativas.

A descrição acima, para fins de explicação, usou nomenclatura específica para fornecer um entendimento completo da invenção. Entretanto, ficará aparente para um indivíduo versado na técnica que detalhes específicos não são exigidos a fim de exercitar a invenção. As descrições acima de modalidades específicas da presente invenção são apresentadas com o objetivo de ilustração e descrição. As mesmas não são entendidas como sendo completas ou como limitantes da invenção precisamente às formas reveladas. Muitas modificações e variações são possíveis em vista dos ensinamentos acima. As modalidades são mostradas e descritas a fim de melhor explicar os princípios da invenção e suas aplicações práticas, para deste modo permitir que outros indivíduos versados na técnica utilizem melhor a invenção e várias modalidades com várias modificações tanto quanto sejam adequadas para o uso particular contemplado. É entendido que o escopo da invenção é definido pelas reivindicações a seguir e seus equivalentes.

REIVINDICAÇÕES

1. Componente de serviço de conceito (408) de um sistema de serviço de busca de conteúdo para buscar um item de conteúdo tendo uma trilha de áudio, o componente de serviço de conceito (408) **caracterizado pelo fato de que** compreende:

um processador de hardware configurado para:

receber, como entrada, um ID de conteúdo (1202) e consulta de busca (1203), em que o ID de conteúdo identifica unicamente o item de conteúdo;

usar o ID de conteúdo (1202) para recuperar um ID de conteúdo (1204), ontologia, vocabulário, e uma transcrição, em que

o ID de categoria refere-se a uma matéria subjetiva do item de conteúdo, e

a transcrição inclui uma renderização textual da trilha de áudio; ;

receber uma consulta de busca (1203) e corrigir e normalizar linguisticamente termos e/ou frases dentro da consulta de busca (1203); e

usar os termos e frases normalizados linguisticamente para processar a transcrição para atribuir pontuações baseadas em ontologia aos termos e/ou frases na transcrição; e

uma memória acoplada ao processador.

2. Componente de serviço de conceito, de acordo com a reivindicação 1, **caracterizado pelo fato de que** o componente de aquisição de recurso solicita um ID de conteúdo (1204) e indicação de data ou data/hora que correspondem ao ID de conteúdo (1202) recebido a partir do componente de armazenamento de dados de mídia do sistema de serviço de busca de conteúdo.

3. Componente de serviço de conceito, de acordo com a reivindicação 2, **caracterizado pelo fato de que** o componente de aquisição de recurso solicita uma ontologia e vocabulário a partir de um componente de armazenamento de ontologia do sistema de serviço de busca de conteúdo,

usando o ID de conteúdo (1204) e indicação de data ou data/hora e uma transcrição a partir de um componente de armazenamento de conteúdo extraído usando o ID de conteúdo (1202) recebido.

4. Componente de serviço de conceito, de acordo com a reivindicação 1, **caracterizado pelo fato de que** o processador de consulta de busca:

aplica regras de linguagem e rotinas baseadas em dicionário aos termos e/ou frases dentro da consulta de busca para corrigir grafias ou quaisquer termos grafados incorretamente na consulta de busca (1203);

aplica rotinas de linguagem para normalizar os termos e/ou frases dentro da consulta de busca (1203) recebida através de troca de formas de plural para formas de singular correspondentes e substituição de termos derivados por formas raiz dos termos derivados; e

filtra dos termos de consulta de busca (1203) aqueles que não ocorrem no vocabulário recebido.

5. Componente de serviço de conceito, de acordo com a reivindicação 1, **caracterizado pelo fato de que** o pontuador de transcrição:

prepara uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente de cada consulta de busca (1203); e

para cada termo e/ou frase na transcrição, associa uma pontuação com o termo e/ou frase baseado nas métricas de co-ocorrência nas listas preparadas de pares de termo/métrica de ontologia.

6. Componente de serviço de conceito, de acordo com a reivindicação 5, **caracterizado pelo fato de que** o pontuador de transcrição prepara uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente de cada consulta de busca (1203) através de:

identificar cada entrada na ontologia que inclui o termo e/ou frase emparelhado com um segundo termo; e

para cada entrada identificada:

computar uma métrica de co-ocorrência como uma

combinação de valores de co-ocorrência na entrada identificada, e

adicionar uma entrada a lista que inclui o segundo termo e a métrica de co-ocorrência computada; e

adicionar uma entrada a lista que inclui o termo e uma métrica de co-ocorrência de termo idêntico.

7. Componente de serviço de conceito, de acordo com a reivindicação 5, **caracterizado pelo fato de que** o pontuador de transcrição, para cada termo e/ou frase na transcrição, associa uma pontuação com o termo e/ou a frase baseada nas métricas de co-ocorrência nas listas preparadas de pares de termo/métrica de ontologia através de:

identificar cada entrada em cada lista de pares de termo/métrica de ontologia na qual a ontologia que inclui o termo e/ou frase considerado correntemente;

quando duas ou mais entradas são identificadas, adicionar as métricas de co-ocorrência das entradas identificadas juntas e computar uma pontuação a partir da soma;

quando uma entrada é identificada, usar a métrica de co-ocorrência na entrada identificada como pontuação; e

associar a pontuação com o termo e/ou frase considerado correntemente.

8. Método para buscar e identificar pontos em um item de conteúdo de mídia transcrita relacionado com uma consulta de busca (1203), **caracterizado pelo fato de que** compreende as etapas de:

receber, como entrada, um ID de conteúdo (1202) e consulta de busca (1203), em que o ID de conteúdo identifica unicamente um item de conteúdo particular;

usar o ID de conteúdo (1202) para recuperar um ID de conteúdo (1204), ontologia, vocabulário, e uma transcrição, em que

o ID de categoria relaciona-se a uma matéria subjetiva do item de conteúdo; e

a transcrição inclui uma renderização textual de uma trilha de áudio do item de conteúdo;

corrigir e normalizar linguisticamente termos e/ou frases dentro da consulta de busca (1203); e

usar os termos e frases normalizados linguisticamente para processar a transcrição para atribuir pontuações baseadas em ontologia aos termos e/ou frases na transcrição.

9. Método, de acordo com a reivindicação 8, **caracterizado pelo fato de que** ainda compreende a etapa de solicitar um ID de conteúdo (1204) e indicação de data ou data/hora que corresponde ao ID de conteúdo (1202) recebido a partir de um componente de armazenamento de dados de mídia de um sistema de serviço de busca de conteúdo.

10. Método, de acordo com a reivindicação 9, **caracterizado pelo fato de que** ainda compreende as etapas de solicitar uma ontologia e vocabulário a partir de um componente de armazenamento de ontologia do sistema de serviço de busca de conteúdo, usar o ID de conteúdo (1204) e indicação de data ou data/hora e solicitar uma transcrição a partir de um componente de armazenamento de conteúdo extraído do sistema de serviço de busca de conteúdo usando o ID de conteúdo (1202).

11. Método, de acordo com a reivindicação 8, **caracterizado pelo fato de que** as etapas de corrigir e normalizar linguisticamente termos e/ou frases dentro da consulta de busca (1203) adicionalmente compreendem as etapas de:

aplicar regras de linguagem e rotinas baseadas em dicionário aos termos e/ou frases dentro da consulta de busca (1203) para corrigir grafias ou quaisquer termos grafados incorretamente na consulta de busca (1203);

aplicar rotinas de linguagem para normalizar os termos e/ou frases dentro da consulta de busca (1203) recebida através de trocar formas de plural para formas de singular correspondentes e substituir termos derivados por formas raiz dos termos derivados; e

filtrar dos termos de consulta de busca (1203) aqueles que não ocorrem no vocabulário recebido.

12. Método, de acordo com a reivindicação 8, **caracterizado pe-**

lo fato de que a etapa de processar a transcrição para atribuir pontuações baseadas em ontologia aos termos e/ou frases adicionalmente compreende as etapas:

preparar uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente de cada consulta de busca (1203); e

para cada termo e/ou frase na transcrição, associar uma pontuação com o termo e/ou frase baseado nas métricas de co-ocorrência nas listas preparadas de pares de termo/métrica de ontologia.

13. Método, de acordo com a reivindicação 12, **caracterizado pelo fato de que** a etapa de preparar uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente da consulta de busca (1203) adicionalmente compreende as etapas de:

identificar cada entrada na ontologia que inclui o termo e/ou frase emparelhado com um segundo termo; e

para cada entrada identificada:

computar uma métrica de co-ocorrência como uma combinação de valores de co-ocorrência na entrada identificada, e

adicionar uma entrada a lista que inclui o segundo termo e a métrica de co-ocorrência computada; e

adicionar uma entrada a lista que inclui o termo e uma métrica de co-ocorrência de termo idêntico.

14. Método, de acordo com a reivindicação 12, **caracterizado pelo fato de que** ainda compreende, para cada termo e/ou frase na transcrição considerado correntemente, a etapa de associar uma pontuação com o termo e/ou a frase baseada nas métricas de co-ocorrência nas listas preparadas de pares de termo/métrica de ontologia através das etapas de:

identificar cada entrada em cada lista de pares de termo/métrica de ontologia na qual a ontologia que inclui o termo e/ou frase considerado correntemente;

quando duas ou mais entradas são identificadas, adicionar as

métricas de co-ocorrência das entradas identificadas juntas e computar uma pontuação a partir da soma;

quando uma entrada é identificada, usar a métrica de co-ocorrência na entrada identificada como pontuação; e

associar a pontuação com o termo e/ou frase considerado correntemente.

15. Meio de armazenamento legível por computador tendo um conjunto de instruções para buscar e identificar pontos em um item de conteúdo de mídia transcrita relacionado com uma consulta de busca (1203), **caracterizado pelo fato de que** as instruções, quando executadas por pelo menos um computador, faz com que pelo menos um computador:

receba, como entrada, um ID de conteúdo (1202) e consulta de busca (1203), em que o ID de conteúdo identifica unicamente um item de conteúdo particular;

use o ID de conteúdo (1202) para recuperar um ID de conteúdo (1204), ontologia, vocabulário, e uma transcrição, em que

o ID de categoria relaciona-se a uma matéria subjetiva do item de conteúdo; e

a transcrição inclui uma renderização textual de uma trilha de áudio do item de conteúdo;

corrija e normalize linguisticamente termos e/ou frases dentro da consulta de busca (1203); e

use os termos e frases normalizados linguisticamente para processar a transcrição para atribuir pontuações baseadas em ontologia aos termos e/ou frases na transcrição.

16. Meio de armazenamento legível por computador, de acordo com a reivindicação 15, **caracterizado pelo fato de que** compreende instruções adicionais para solicitar um ID de conteúdo (1204) e indicação de data ou data/hora que corresponde ao ID de conteúdo (1202) recebido a partir de um componente de armazenamento de dados de mídia de um sistema de serviço de busca de conteúdo.

17. Meio de armazenamento legível por computador, de acordo

com a reivindicação 16, **caracterizado pelo fato de que** compreende instruções adicionais para solicitar uma ontologia e vocabulário a partir de um componente de armazenamento de ontologia do sistema de serviço de busca de conteúdo, usar o ID de conteúdo (1204) e indicação de data ou data/hora e solicitar uma transcrição a partir de um componente de armazenamento de conteúdo extraído do sistema de serviço de busca de conteúdo usando o ID de conteúdo (1202).

18. Meio de armazenamento legível por computador, de acordo com a reivindicação 15, **caracterizado pelo fato de que** as instruções para de corrigir e normalizar linguisticamente termos e/ou frases dentro da consulta de busca (1203) ainda compreendem as etapas de:

aplicar regras de linguagem e rotinas baseadas em dicionário aos termos e/ou frases dentro da consulta de busca (1203) para corrigir grafias ou quaisquer termos grafados incorretamente na consulta de busca (1203);

aplicar rotinas de linguagem para normalizar os termos e/ou frases dentro da consulta de busca (1203) recebida através de trocar formas de plural para formas de singular correspondentes e substituir termos derivados por formas raiz dos termos derivados; e

filtrar dos termos de consulta de busca (1203) aqueles que não ocorrem no vocabulário recebido.

19. Meio de armazenamento legível por computador, de acordo com a reivindicação 15, **caracterizado pelo fato de que** as instruções adicionais para processar a transcrição para atribuir pontuações baseadas em ontologia aos termos e/ou frases adicionalmente compreende as etapas de:

preparar uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente de cada consulta de busca (1203); e

para cada termo e/ou frase na transcrição, associar uma pontuação com o termo e/ou frase baseado nas métricas de co-ocorrência nas listas preparadas de pares de termo/métrica de ontologia.

20. Meio de armazenamento legível por computador, de acordo

com a reivindicação 19, **caracterizado pelo fato de que** as instruções adicionais para preparar uma lista de pares de termo/métrica de ontologia para cada termo e/ou frase nos termos e/ou frases normalizados linguisticamente da consulta de busca (1203) adicionalmente compreende as etapas de:

identificar cada entrada na ontologia que inclui o termo e/ou frase emparelhado com um segundo termo; e

para cada entrada identificada:

computar uma métrica de co-ocorrência como uma combinação de valores de co-ocorrência na entrada identificada, e

adicionar uma entrada a lista que inclui o segundo termo e a métrica de co-ocorrência computada; e

adicionar uma entrada a lista que inclui o termo e uma métrica de co-ocorrência de termo idêntico.

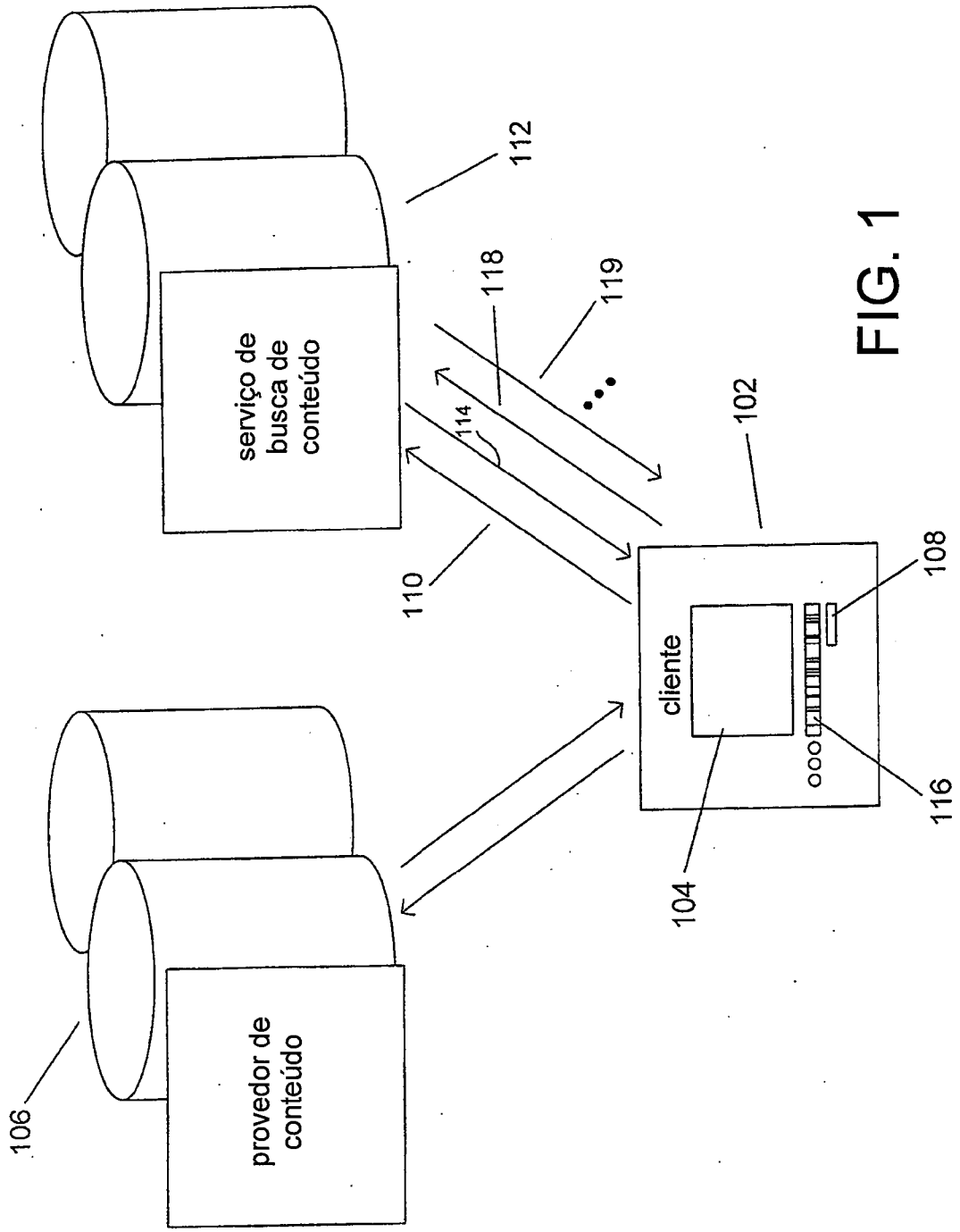

serviço de busca de conteúdo
112
118
119
114
102
110
cliente
108
116
104
provedor de conteúdo
106

FIG. 1

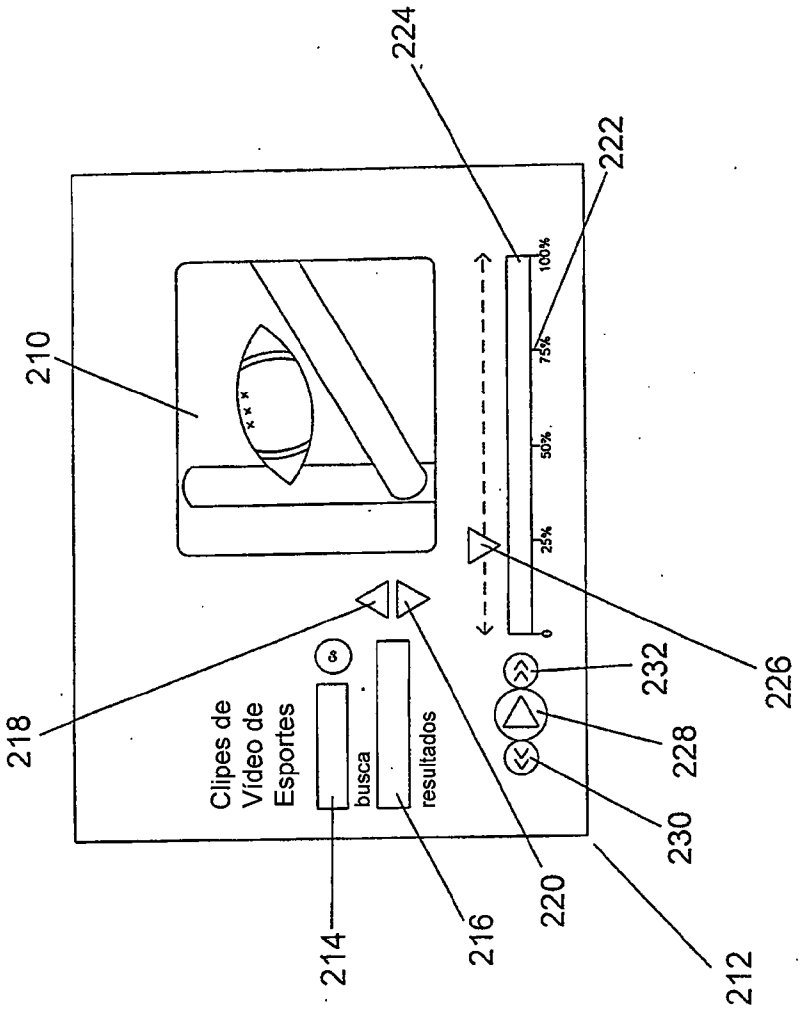

224
222
210
100%
75%
50%
25%
0
218
226
232
228
230
Clipes de Vídeo de Esportes
busca
resultados
214
216
220
212

FIG. 2

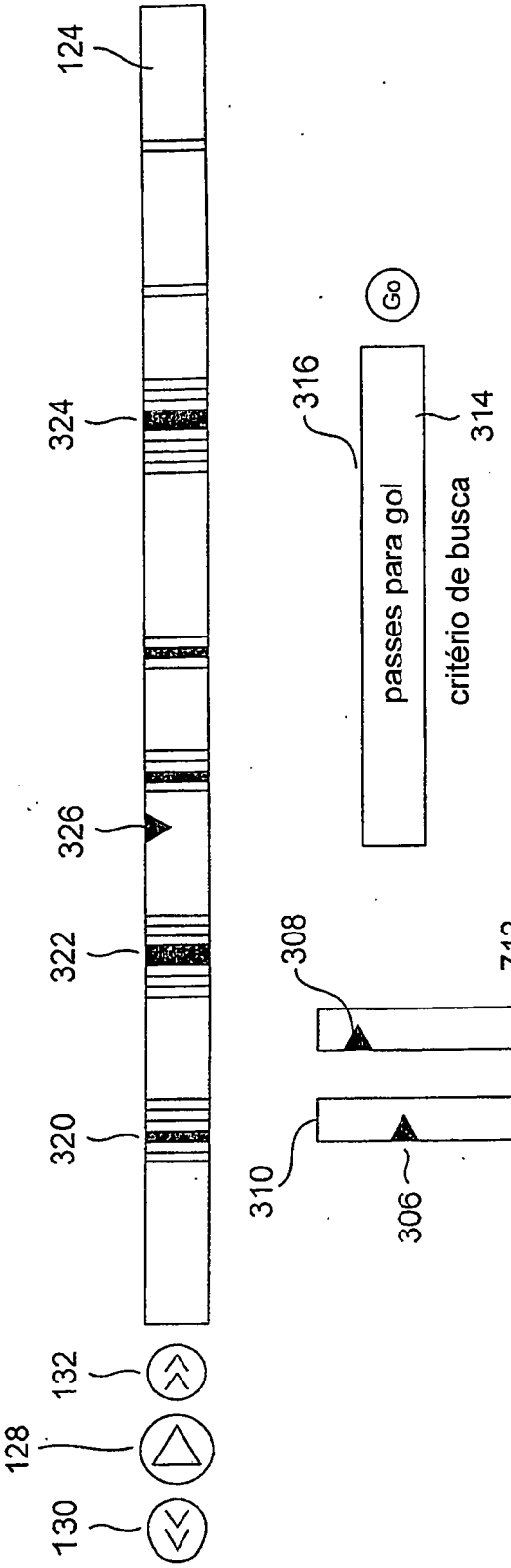

124
324
326
322
320
132
128
130
316
Go
passes para gol
critério de busca
314
308
712
304
suavização
310
306
limite
302

FIG. 3

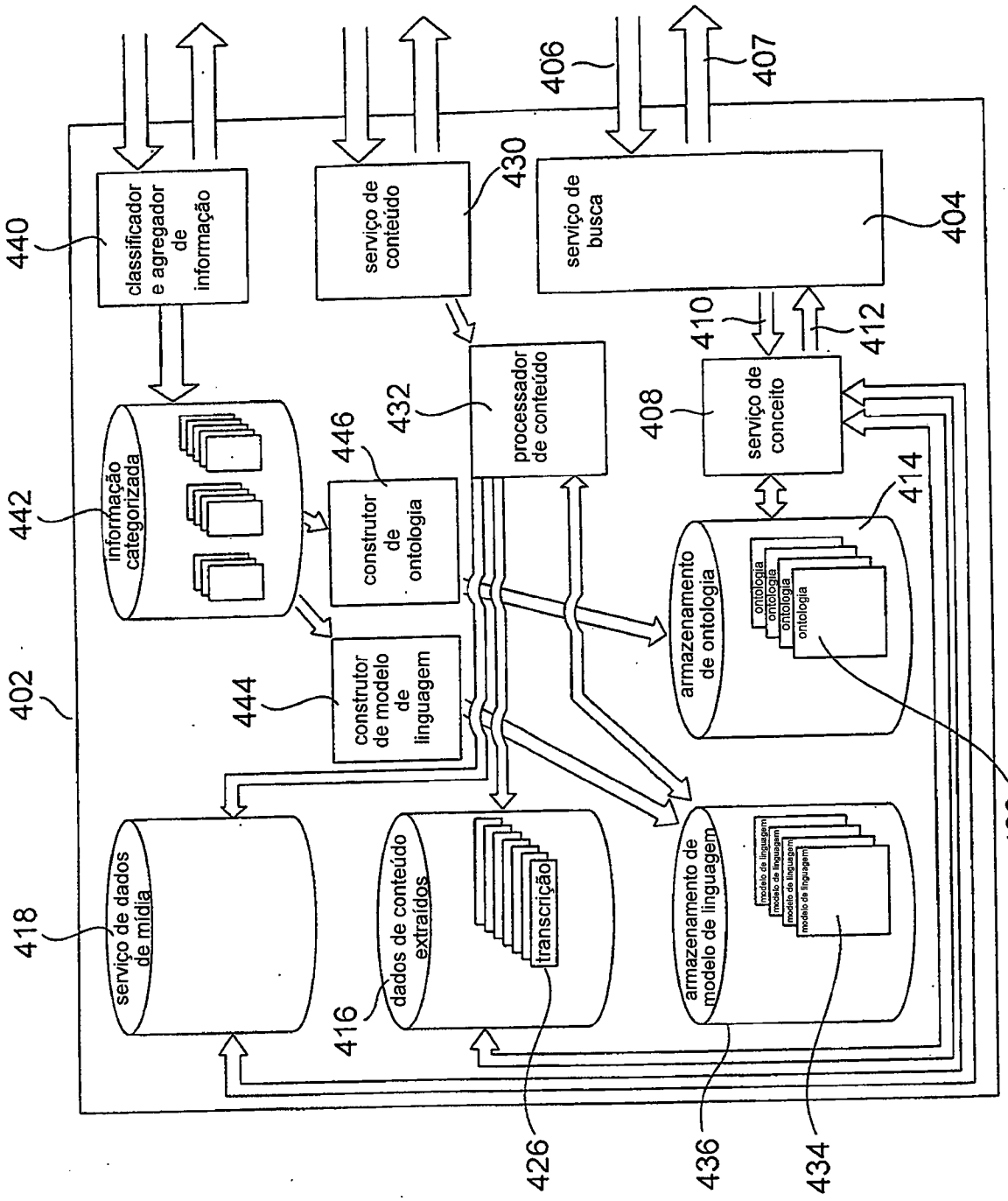

402
404
406
407
408
410
412
414
416
418
422
426
430
432
434
436
440
442
444
446
classificador e agregador de informação
serviço de conteúdo
serviço de busca
processador de conteúdo
serviço de conceito
informação categorizada
construtor de ontologia
construtor de modelo de linguagem
armazenamento de ontologia
ontologia
ontologia
ontologia
ontologia
serviço de dados de mídia
dados de conteúdo extraídos
transcrição
armazenamento de modelo de linguagem
modelo de linguagem
modelo de linguagem
modelo de linguagem
modelo de linguagem

FIG. 4

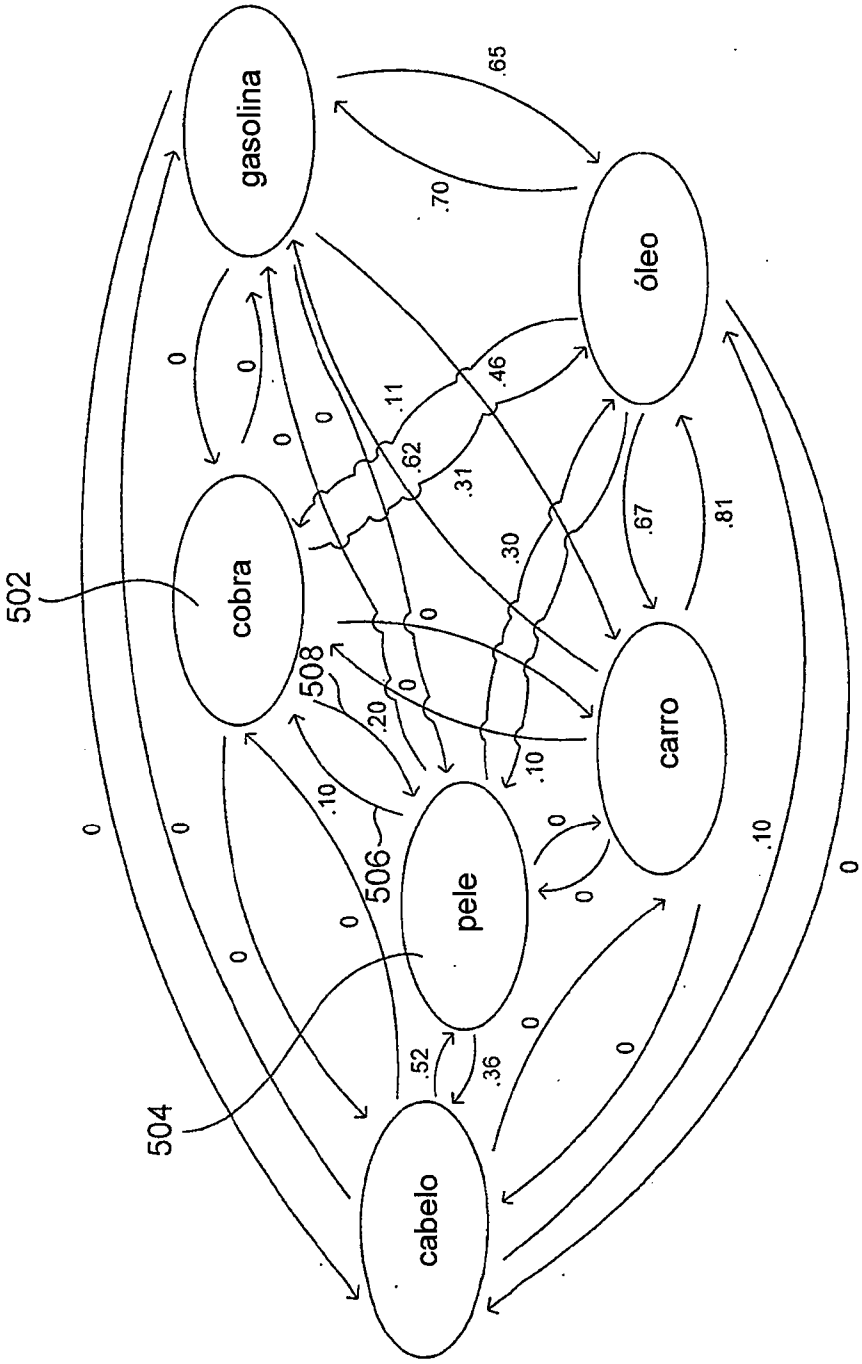

gasolina
óleo
cobra
carro
pele
cabelo
502
504
506
508
.65
.70
.11
.46
.62
.31
.30
.67
.81
.20
.10
.52
.36

FIG. 5

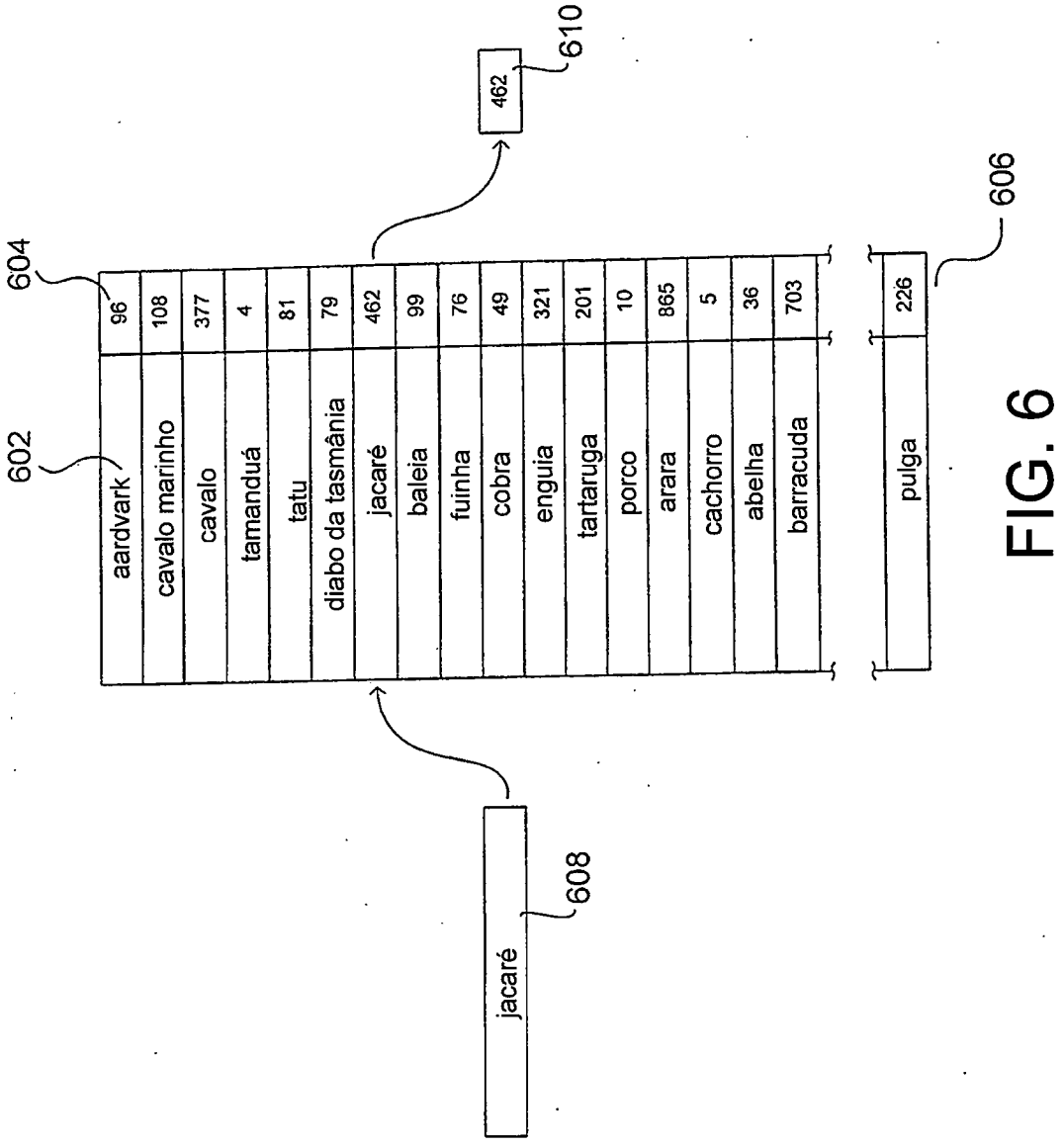

602
604
606
608
610
aardvark
cavalo marinho
cavalo
tamanduá
tatu
diabo da tasmânia
jacaré
baleia
fuinha
cobra
enguia
tartaruga
porco
arara
cachorro
abelha
barracuda
96
108
377
4
81
79
462
99
76
49
321
201
10
865
5
36
703
pulga
226
462
jacaré

FIG. 6

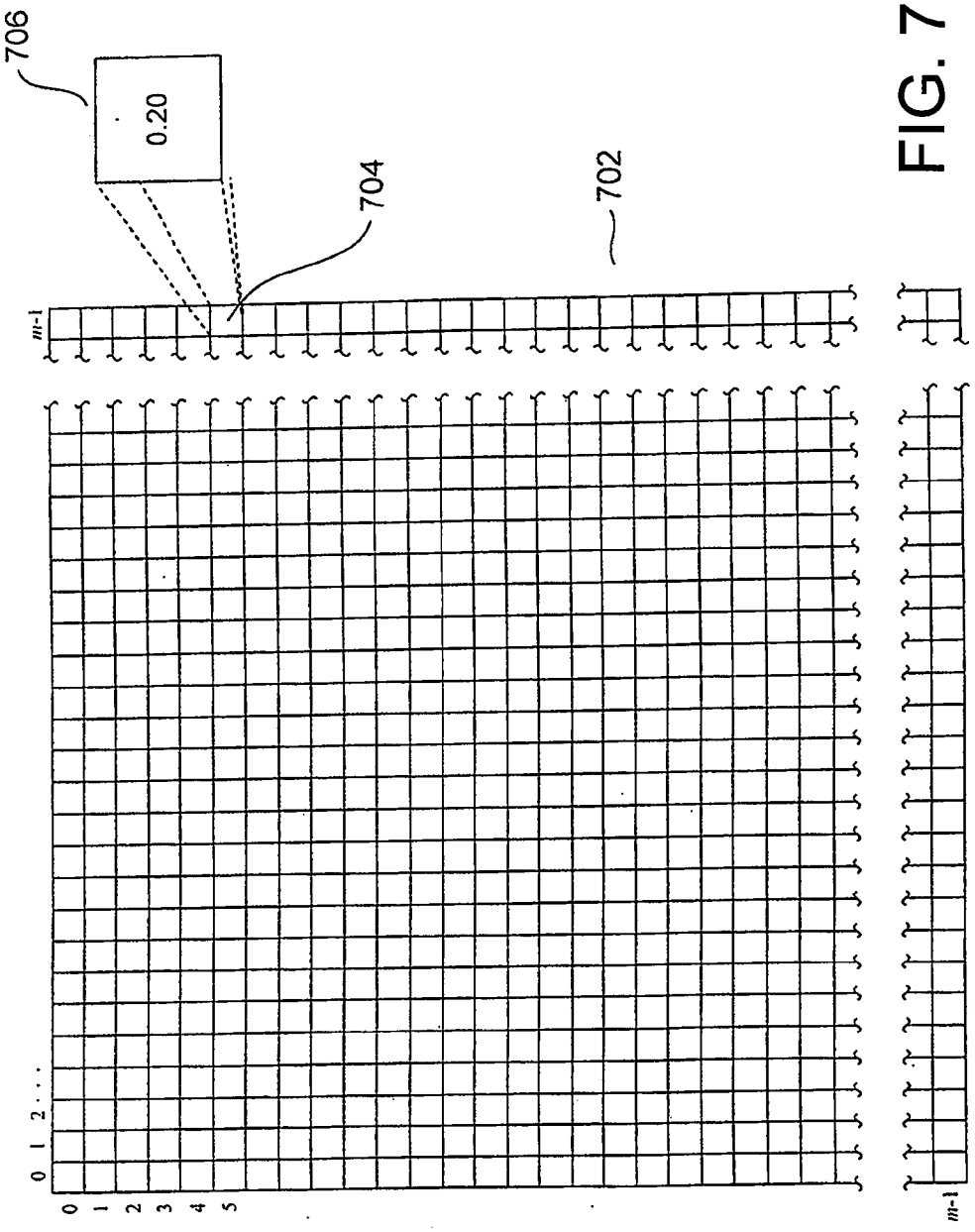
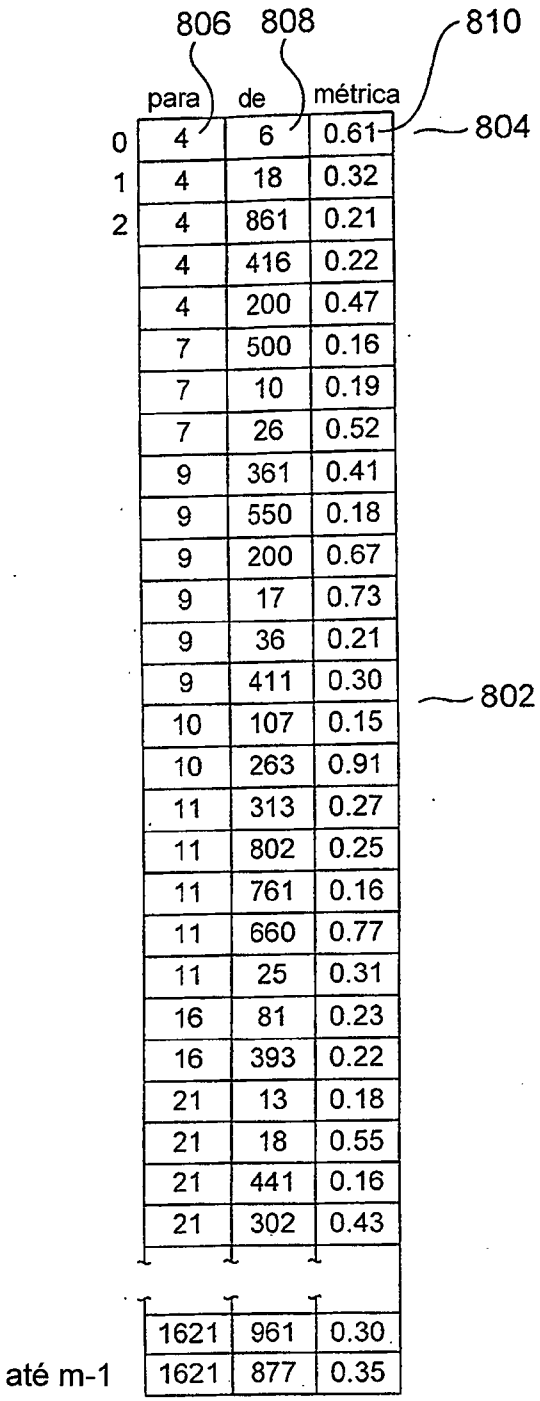


FIG. 7



	para	de	métrica
0	4	6	0.61
1	4	18	0.32
2	4	861	0.21
	4	416	0.22
	4	200	0.47
	7	500	0.16
	7	10	0.19
	7	26	0.52
	9	361	0.41
	9	550	0.18
	9	200	0.67
	9	17	0.73
	9	36	0.21
	9	411	0.30
	10	107	0.15
	10	263	0.91
	11	313	0.27
	11	802	0.25
	11	761	0.16
	11	660	0.77
	11	25	0.31
	16	81	0.23
	16	393	0.22
	21	13	0.18
	21	18	0.55
	21	441	0.16
	21	302	0.43
	1621	961	0.30
até m-1	1621	877	0.35

FIG. 8

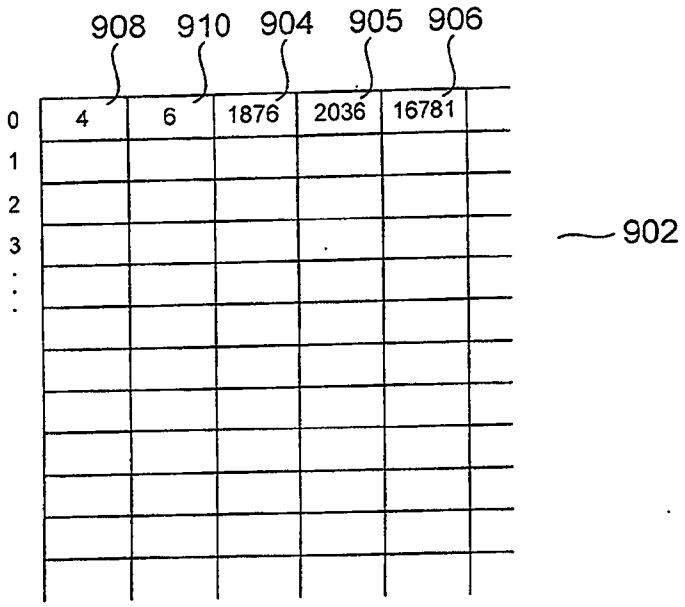


FIG. 9

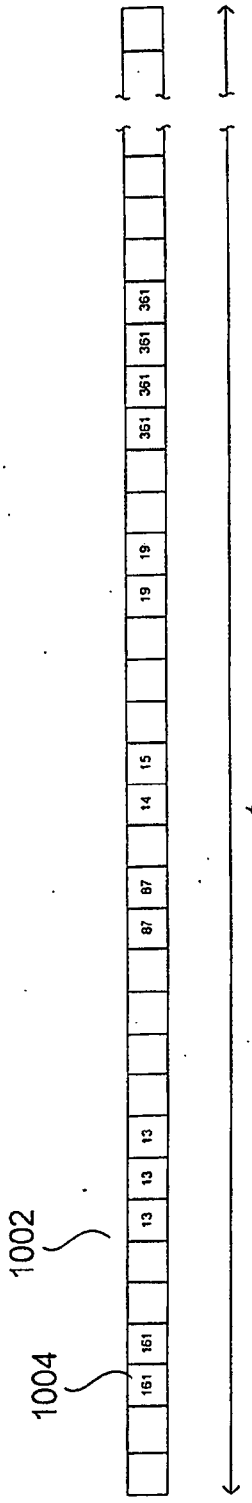


FIG. 10

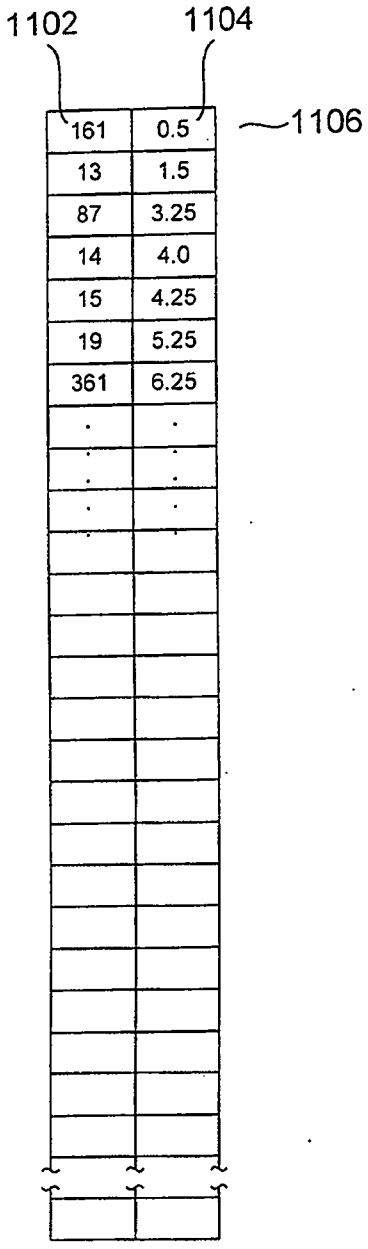


FIG. 11

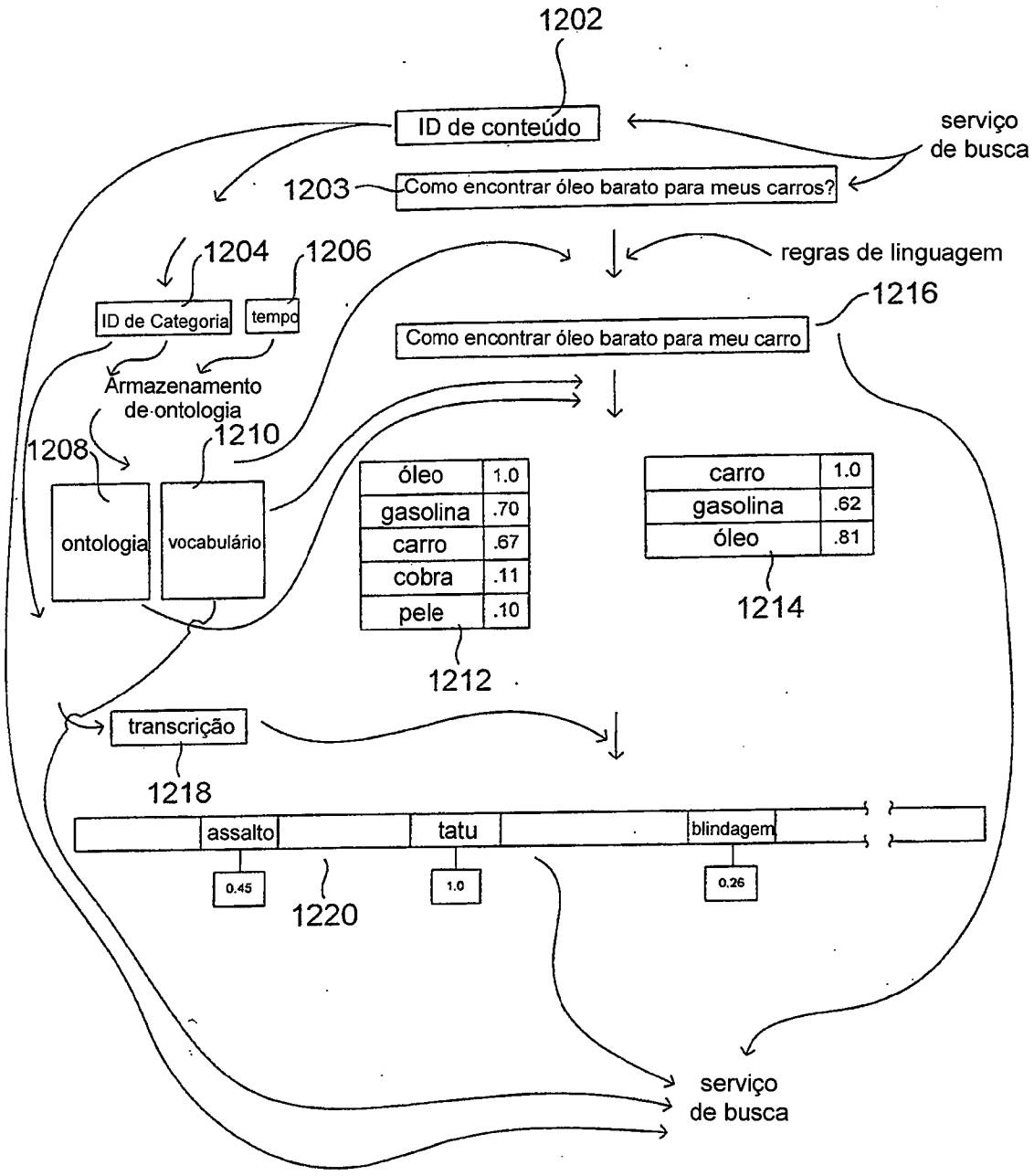

1202
ID de conteúdo
serviço de busca
1203
Como encontrar óleo barato para meus carros?
1204
1206
regras de linguagem
ID de Categoria
tempo
1216
Como encontrar óleo barato para meu carro
Armazenamento de ontologia
1208
1210
ontologia
vocabulário
óleo 1.0
gasolina .70
carro .67
cobra .11
pele .10
1212
carro 1.0
gasolina .62
óleo .81
1214
transcrição
1218
assalto
tatu
blindagem
0.45
1.0
0.26
1220
serviço de busca

FIG. 12

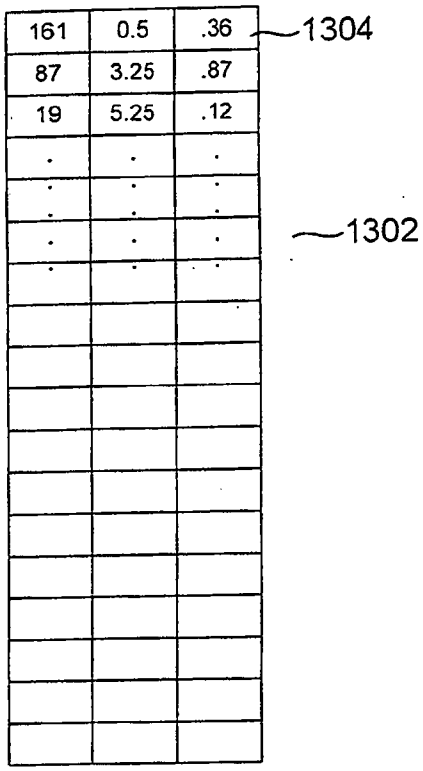


FIG. 13

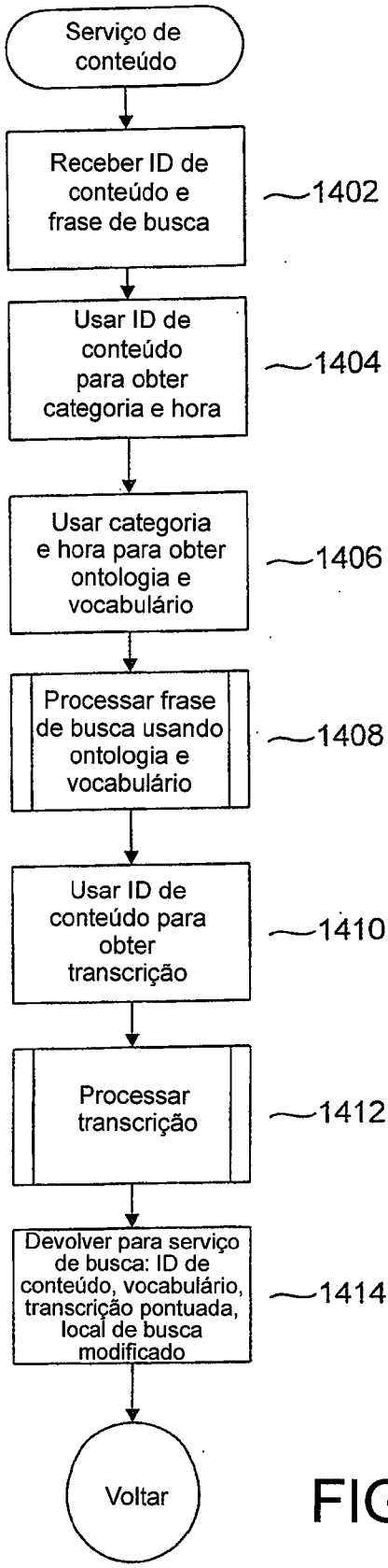

Serviço de conteúdo
Receber ID de conteúdo e frase de busca
1402
Usar ID de conteúdo para obter categoria e hora
1404
Usar categoria e hora para obter ontologia e vocabulário
1406
Processar frase de busca usando ontologia e vocabulário
1408
Usar ID de conteúdo para obter transcrição
1410
Processar transcrição
1412
Devolver para serviço de busca: ID de conteúdo, vocabulário, transcrição pontuada, local de busca modificado
1414
Voltar

FIG. 14

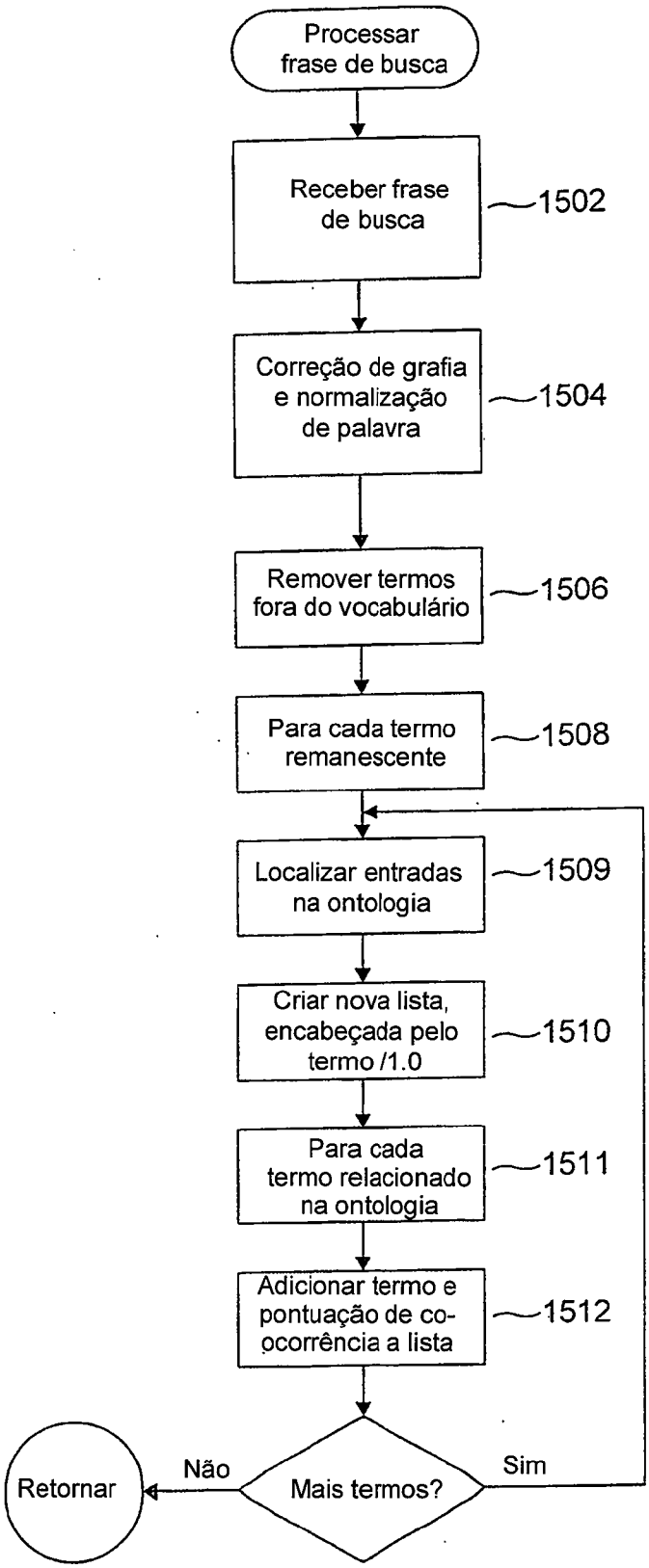

Processar frase de busca
Receber frase de busca
1502
Correção de grafia e normalização de palavra
1504
Remover termos fora do vocabulário
1506
Para cada termo remanescente
1508
Localizar entradas na ontologia
1509
Criar nova lista, encabeçada pelo termo /1.0
1510
Para cada termo relacionado na ontologia
1511
Adicionar termo e pontuação de co-ocorrência a lista
1512
Mais termos?
Não
Sim
Retornar

FIG. 15

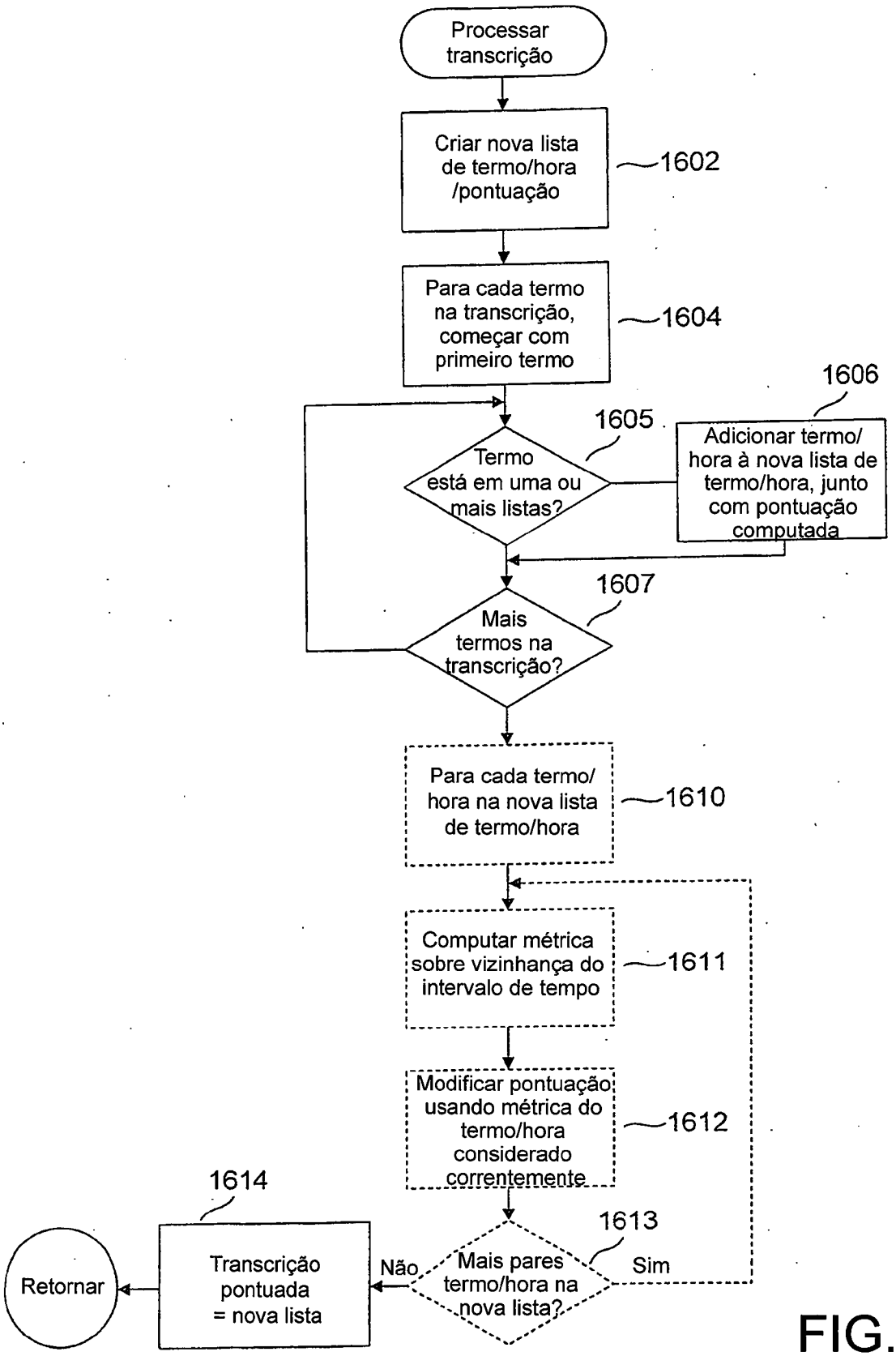

Processar transcrição
Criar nova lista de termo/hora /pontuação
1602
Para cada termo na transcrição, começar com primeiro termo
1604
1606
1605
Termo está em uma ou mais listas?
Adicionar termo/ hora à nova lista de termo/hora, junto com pontuação computada
1607
Mais termos na transcrição?
Para cada termo/ hora na nova lista de termo/hora
1610
Computar métrica sobre vizinhança do intervalo de tempo
1611
Modificar pontuação usando métrica do termo/hora considerado correntemente
1612
1614
1613
Retornar
Transcrição pontuada = nova lista
Não
Mais pares termo/hora na nova lista?
Sim

FIG. 16

RESUMO

Patente de Invenção: **"COMPONENTE DE SERVIÇO DE CONCEITO DE UM SISTEMA DE BUSCA DE CONTEÚDO, MÉTODO PARA BUSCAR E IDENTIFICAR PONTOS EM UM ITEM DE CONTEÚDO DE MÍDIA E MEIO DE ARMAZENAMENTO LEGÍVEL POR COMPUTADOR TENDO UM CONJUNTO DE INSTRUÇÕES PARA REALIZAR O REFERIDO MÉTODO".**

A presente invenção refere-se a várias modalidades que incluem componentes de serviço de conceito dos sistemas de serviço de busca de conteúdo que empregam ontologias e vocabulários preparados para categorias particulares de conteúdo em tempos particulares a fim de pontuar transcrições preparadas a partir de itens de conteúdo para permitir um componente de serviço de busca de um sistema de serviço de busca de conteúdo a atribuir estimativas de afinidade de partes de um item de conteúdo com o critério de busca, a fim de apresentar resultados de busca para os clientes do sistema de serviço de busca de conteúdo. O componente de serviço de conceito processa uma solicitação de busca para gerar listas de termos relacionados, e em seguida emprega as listas de termos relacionados para processar transcrições a fim de pontuar transcrições baseado na informação contida nas ontologias.