



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

(21) **PI0708330-0 A2**



* B R P I 0 7 0 8 3 3 0 A 2 *

(22) Data de Depósito: 27/02/2007
(43) Data da Publicação: 24/05/2011
(RPI 2107)

(51) *Int.Cl.:*
G06F 17/30 2006.01
G06F 17/21 2006.01

(54) Título: **TREINAMENTO DE UMA FUNÇÃO DE CLASSIFICAÇÃO USANDO RELEVÂNCIA PROPAGADA DE DOCUMENTOS**

(30) Prioridade Unionista: 27/02/2006 US 11/364.576

(73) Titular(es): Microsoft Corporation

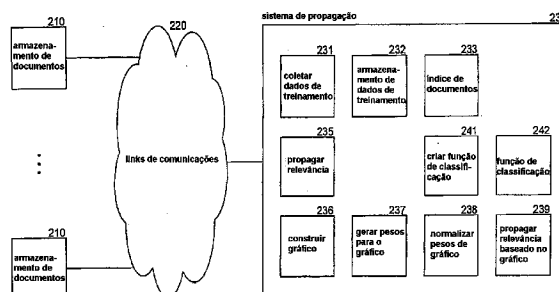
(72) Inventor(es): Jue Wang, Mingjing Li, Wei-Ying Ma, Zhiwei Li

(74) Procurador(es): Nellie Anne Danie-Shores

(86) Pedido Internacional: PCT US2007005141 de 27/02/2007

(87) Publicação Internacional: WO 2007/100844 de 07/09/2007

(57) **Resumo:** TREINAMENTO DE UMA FUNÇÃO DE CLASSIFICAÇÃO USANDO RELEVÂNCIA PROPAGADA DE DOCUMENTOS Um método e sistema para propagar a relevância de documentos rotulados para uma consulta para documentos não rotulados são fornecidos. O sistema de propagação fornece dados de treinamento que inclui consultas, documentos rotulados com sua relevância para as consultas, e documentos não rotulados. O sistema de propagação então calcula a similaridade entre pares de documentos nos dados de treinamento. O sistema de propagação então propaga a relevância dos documentos rotulados para documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos converjam para uma solução. Os dados de treinamento com as relevâncias propagadas podem então ser usados para treinar uma função de classificação.





PI0708330-0

"TREINAMENTO DE UMA FUNÇÃO DE CLASSIFICAÇÃO USANDO RELEVÂNCIA PROPAGADA DE DOCUMENTOS"

ANTECEDENTES DA INVENÇÃO

Muitos serviços de motor de busca, tais como Google e Overture, fornecem pesquisa de informação que é acessível via Internet. Estes serviços de motor de busca permitem que os usuários pesquisem páginas de exibição, tais como páginas da rede, que podem ser de interesse para os usuários. Após um usuário apresentar uma solicitação de busca (isto é, uma consulta) que inclui termos de busca, o serviço de motor de busca identifica as páginas da rede que podem estar relacionadas àqueles termos de busca. Para identificar rapidamente páginas da rede relacionadas, os serviços de motor de busca podem manter um mapeamento de palavras-chaves para as páginas da rede. Este mapeamento pode ser gerado por "arrastar" a rede (isto é, a World Wide Web) para identificar as palavras-chaves de cada página da rede. Para arrastar a rede, um serviço de motor de busca pode usar uma lista de páginas da rede de raiz para identificar todas as páginas da rede que estão acessíveis através daquelas páginas da rede de raiz. As palavras-chaves de qualquer página da rede particular podem ser identificadas usando várias técnicas de recuperação de informação bastante conhecidas, tais como identificar as palavras de um título, as palavras fornecidas nos metadados da página da rede, as palavras que estão destacadas, e assim por diante. O serviço de motor de busca identifica páginas da rede que podem estar relacionadas à solicitação de pesquisa baseada em como as palavras-chaves de uma página da rede se encaixam com as palavras da consulta. O serviço do motor de busca então exhibe para os usuários links para as páginas da rede identificadas em uma ordem que é baseada em uma classificação que pode ser determinada pela sua relevância para a consulta, popularidade, importância, e/ou alguma outra medida.

Três técnicas bastante conhecidas para a classificação de páginas da rede são PageRank, HITS ("Pesquisa de Tópicos induzida por Hiperlink") e DirectHIT. A PageRank é baseada no princípio de que páginas da rede terão links (isto é, "links de saída") para páginas da rede importantes. Deste modo, a importância de uma página da rede é baseada no número e importância de outras páginas da rede que se conectam àquela página da rede (isto é, "links de entrada"). Em uma forma simples, os links entre páginas da rede podem ser representados pela matriz de adjacência A , onde A_{ij} representa o número de links de saída da página da rede i para a página da rede j . A marcação de importância w_j para a página da rede j pode ser representada pela seguinte equação:

$$W_j = \sum_i A_{ij} w_i$$

Esta equação pode ser resolvida pelos cálculos iterativos baseados na seguinte equação:

$$A^T w = w$$

onde w é o vetor de marcações de importância para as páginas da rede e o eigen-vetor principal de A^T .

A técnica HITS é baseada adicionalmente no princípio de que uma página da rede que tem muitos links com outras páginas da rede importantes, pode ela mesma ser impor-
 5 tante. Deste modo, a HITS divide “importância” de páginas da rede em dois atributos rela-
 cionados: “hub” e “autoridade”. “Hub” é medida pela marcação de “autoridade” das páginas
 da rede com a qual uma página da rede se conecta, e “autoridade” é medida pela marcação
 da “hub” das páginas da rede que se conectam a página da rede. Em contraste com a Pa-
 geRank, que calcula a importância de páginas da rede independentemente da consulta, a
 10 HITS calcula a importância baseada nas páginas da rede do resultado e páginas da rede
 que estão relacionadas às páginas da rede do resultado pelos seguintes links de entrada e
 saída. A HITS apresenta uma consulta para um serviço de motor de busca e usa as páginas
 da rede do resultado como o conjunto inicial de páginas da rede. A HITS adiciona ao conjun-
 to àquelas páginas da rede que são as destinações de links de entrada e aquelas páginas
 15 da rede que são as fontes de links de saída das páginas da rede do resultado. A HITS então
 calcula a marcação da hub e da autoridade de cada página da rede usando um algoritmo
 iterativo. As marcações da hub e da autoridade podem ser representadas pelas seguintes
 equações:

$$A(p) = \sum_{q \rightarrow p} h(q) \text{ e } h(p) = \sum_{p \rightarrow q} a(q)$$

onde $a(p)$ representa a marcação de autoridade para a página da rede p e $h(p)$ re-
 apresenta a marcação da hub para a página da rede p . A HITS usa uma matriz de adjacência
 A para representar os links. A matriz de adjacência é representada pela seguinte equação:

$$b_{ij} = \begin{cases} 1 & \text{se a página } i \text{ tiver um link com a página } j, \\ 0 & \text{de outro modo} \end{cases}$$

Os vetores a e h correspondem às marcações de hub e autoridade, respectivamen-
 te, de todas as páginas da rede no conjunto e podem ser representados pelas seguintes
 equações:

$$a = A^T h \text{ e } h = Aa$$

Deste modo, a e h são eigenvetores de matrizes $A^T A$ e AA^T . A HITS também pode
 ser modificada para fator na popularidade de uma página da rede como medido pelo número
 de visitas. Baseado em uma análise de dados através do clique, b_{ij} da matriz de adjacência
 pode ser aumentado sempre que um usuário se move da página da rede i para a página da
 rede j .

A DirectHit classifica páginas da rede com base na história passada do usuário com
 resultados e consultas similares. Por exemplo, se usuários que apresentam consultas simila-
 res tipicamente selecionaram primeiro a terceira página da rede do resultado, então a histó-

ria deste usuário seria uma indicação de que a terceira página da rede deveria ser classificada mais alta. Como um outro exemplo, se usuários que apresentam consultas similares passam tipicamente a maior parte do tempo vendo a quarta página da rede do resultado, então a história deste usuário seria uma indicação de que a quarta página da rede deveria ser classificada mais alta. A DirectHIT é derivada de histórias de usuários de análise de dados através do clique.

Algumas técnicas de classificação usam algoritmos de aprendizagem de máquinas para aprender uma função de classificação de dados de treinamento que inclui consultas, vetores de características representando páginas, e para cada consulta, uma classificação para cada página. Uma função de classificação serve como um mapeamento de características de uma página para sua classificação para uma dada consulta. A aprendizagem de uma função de classificação tem sido considerada por alguns como um problema de regressão para aprender o mapeamento de um vetor de característica para um membro de um conjunto ordenado de classificações numéricas. Algumas técnicas baseadas em regressão tentam fornecer uma marcação de relevância absoluta que pode ser usada para classificar páginas. Uma função de classificação, entretanto, não precisa fornecer uma marcação de relevância absoluta, mas antes precisa somente fornecer uma classificação relativa das páginas. Deste modo, estas técnicas baseadas em regressão resolvem um problema que é mais difícil do que necessário.

Algoritmos de aprendizagem de máquinas para uma função de classificação usam consultas, vetores de características, marcações de relevância rotuladas pelo usuário como dados de treinamento. Para gerar os dados de treinamento, consultas podem ser apresentadas a um motor de busca que gera as páginas do resultado de pesquisa. Os algoritmos então geram os vetores de característica para as páginas e entrada de um usuário das marcações de relevância para cada página. Uma dificuldade com tal abordagem é que um motor de busca pode retornar centenas de páginas como seu resultado de pesquisa. Pode ser bem caro para um usuário rotular todas as páginas de um resultado de pesquisa. Além disso, pode ser difícil para um usuário avaliar precisamente a relevância de um número tão grande de páginas. Embora um usuário pudesse rotular somente uma pequena porção das páginas, a aprendizagem baseada em tal porção pequena pode não fornecer uma função de classificação precisa.

SUMÁRIO

Um método e sistema para propagar a relevância de documentos rotulados para uma consulta em relação à relevância de documentos não rotulados são fornecidos. O sistema de propagação provê dados de treinamento que incluem consultas, documentos rotulados com sua relevância para as consultas e documentos não rotulados. O sistema de propagação então calcula a similaridade entre pares de documentos nos dados de treinamento.

O sistema de propagação então propaga a relevância dos documentos rotulados para documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos converjam para uma solução. Os dados de treinamento com as relevâncias reproduzidas podem então ser usados para treinar uma

5

função de classificação.

Este Sumário é fornecido para apresentar uma seleção de conceitos de uma forma simplificada que são ainda descritos abaixo na Descrição Detalhada. Este sumário não pretende identificar características chaves ou características essenciais do assunto reivindicado, nem pretende ser usado como uma ajuda na determinação do escopo do assunto reivin-

10

BREVE DESCRIÇÃO DOS DESENHOS

A figura 1 é um diagrama que ilustra uma porção de um gráfico de documentos.

A figura 2 é um diagrama de blocos que ilustra componentes do sistema de propagação em uma modalidade.

15

A figura 3 é um diagrama de fluxo que ilustra o processamento do componente de criar função de classificação do sistema de propagação em uma modalidade.

A figura 4 é um diagrama de fluxo que ilustra o processamento do componente de propagar relevância do sistema de propagação em uma modalidade.

20

A figura 5 é um diagrama de fluxo que ilustra o processamento do componente de construir gráfico do sistema de propagação em uma modalidade.

A figura 6 é um diagrama de fluxo que ilustra o processamento do componente de gerar pesos para gráfico do sistema de propagação em uma modalidade.

A figura 7 é um diagrama de fluxo que ilustra o processamento do componente de normalizar pesos de gráfico do sistema de propagação em uma modalidade.

25

A figura 8 é um diagrama de fluxo que ilustra o processamento do componente de propagar relevância baseada em gráfico do sistema de propagação em uma modalidade.

DESCRIÇÃO DETALHADA

Um método e sistema para propagar relevância de documentos rotulados para uma consulta a documentos não rotulados são fornecidos. Em uma modalidade, o sistema de propagação fornece dados de treinamento que incluem consultas, documentos (representados por vetores de características) rotulados com sua relevância para as consultas, e documentos não rotulados. Por exemplo, o sistema de propagação pode apresentar uma consulta a um motor de busca e usar o resultado da pesquisa como os documentos (por exemplo, páginas da rede). O sistema de propagação pode então induzir um usuário a rotular alguns dos documentos do resultado da pesquisa baseado em sua relevância para a consulta. O sistema de propagação então calcula a similaridade entre pares de documentos nos dados de treinamento. Por exemplo, o sistema de propagação pode representar cada documento

30

35

por um vetor de características e pode calcular a similaridade entre documentos baseado na distância Euclidiana no espaço de características ou baseado em uma métrica de similaridade de co-seno. O sistema de propagação então propaga a relevância dos documentos rotulados para documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos convirjam em uma solução. Os dados de treinamento com as relevâncias propagadas podem então ser usados para treinar uma função de classificação. Deste modo, o sistema de propagação pode automaticamente aumentar dados de treinamento com dados de treinamento adicionais baseados nas similaridades entre documentos.

Em uma modalidade, o sistema de propagação representa os documentos usando um gráfico de documento com cada nó representando um documento e cada borda representando similaridade entre os documentos representados pelos nós conectados. O sistema de propagação pode representar o gráfico como uma matriz quadrada com uma linha e coluna para cada documento no qual cada valor não-zero indica uma borda entre o nó da linha e o nó da coluna. O sistema de propagação pode definir bordas para o gráfico usando várias técnicas. Por exemplo, o sistema de propagação pode considerar que o gráfico esteja totalmente conectado em cujo caso cada nó tem uma borda para nós alternados. Como um outro exemplo, o sistema de propagação pode considerar os nós a serem conectados por meio de uma árvore de cobertura mínima. Em uma modalidade, o sistema de propagação considera os nós a serem conectados usando um algoritmo vizinho mais próximo de k . Em particular, o sistema de propagação identifica os vizinhos mais próximos de k para cada nó e adiciona uma borda daquele nó para cada um dos seus vizinhos mais próximos de k . O sistema de propagação então calcula pesos para as bordas, baseado na similaridade entre os documentos representados pelas bordas conectadas. O sistema de propagação pode usar várias técnicas para determinar a similaridade entre documentos. Em uma modalidade, o sistema de propagação usa uma métrica de distância Euclidiana baseada na representação do vetor de características dos documentos em um espaço de características. O sistema de propagação armazena a similaridade como os valores da matriz quadrada resultando em uma matriz de afinidade ou similaridade. O sistema de propagação pode também normalizar a matriz de similaridade. O sistema de propagação pode também definir os valores diagonais para 0 para evitar auto-reforço durante a propagação de relevância.

Após gerar a matriz de similaridade, o sistema de propagação propaga a relevância dos documentos rotulados para os documentos não rotulados usando um algoritmo de propagação baseado em classificação de tubulação. Um algoritmo baseado em classificação de tubulação é descrito em He, J., Li, M., Zhang, H.J, e outros, "Manifold-Ranking Based Image Retrieval", Proc. da 12ª Conferência Internacional ACM Anual sobre Multimídia, 2004. O sistema de propagação define inicialmente a relevância dos documentos rotulados para a

marcação de relevância fornecida pelo usuário e a relevância dos documentos não rotulados para 0. O sistema de propagação então dispersa a relevância dos documentos rotulados para os seus documentos não rotulados conectados reduzindo a fatores na similaridade como indicado pela matriz de similaridade. O sistema de propagação dispersa iterativamente a
 5 marcação de relevância até que as marcações de relevância converjam em uma solução. As marcações de relevância resultantes dos documentos não rotulados estarão em proporção à probabilidade que eles são relevantes para a mesma consulta que os documentos rotulados. Um documento não rotulado que é muito similar a muitos documentos rotulados com marcações de alta relevância terá deste modo uma marcação de alta relevância. De modo inverso,
 10 um documento não rotulado que não é muito similar a quaisquer documentos rotulados terá uma marcação de baixa relevância.

O sistema de propagação pode representar similaridade usando um núcleo Laplace, que pode ser representado pela seguinte equação:

$$k_L(x_i, x_j) = \prod_{l=1}^t \frac{1}{2\sigma_l} \exp(-|x_{il} - x_{jl}|/\sigma_l) \quad (1)$$

15 Onde x_{ij} e x_{ji} representam a dimensão l de x_i e x_j respectivamente, i representa a dimensionalidade do espaço de características, e σ_l representa um parâmetro positivo que reflete os pesos de dimensões diferentes no cálculo da similaridade. Deste modo, o sistema de propagação representa o peso das bordas pela seguinte equação:

$$W_{ij} = k_L(x_i, x_j) = \prod_{l=1}^t \exp(-|x_{il} - x_{jl}|/\sigma_l) \quad (2)$$

20 Onde W_{ij} representa a similaridade entre os documentos i e j . O sistema de propagação pode omitir o coeficiente constante $1/2\sigma_l$ já que seu efeito sobre a matriz de similaridade W será neutralizado pela normalização da matriz. O sistema de propagação normaliza a matriz de similaridade como representado pela seguinte equação:

$$S = D^{-1/2} W D^{-1/2} \quad (3)$$

25 onde S representa a matriz de similaridade normalizada e D representa uma matriz diagonal onde (i,i) é igual à soma da i linha da matriz de similaridade W . A normalização normaliza as similaridades para serem relativas à similaridade dos documentos conectados.

O sistema de propagação pode representar cada documento como um vetor x de características de dimensão t que forma um ponto no espaço Euclidiano. Para uma consulta, o sistema de propagação recebe o conjunto de resultados de documentos

$$\mathcal{X} = \{x_{i1}, x_{i2}, \dots, x_{im}, x_{u1}, x_{u2}, \dots, x_{un}\} \subset \mathbb{R}^t.$$

Os primeiros pontos m (no espaço de características) representam documentos rotulados pelo usuário, e os últimos pontos n (no espaço de características) representam documentos não rotulados. O sistema de propagação também recebe um vetor de rótulos correspondente

$$\gamma = [y_{i1}, y_{i2}, \dots, y_{im}, 0, 0, \dots, 0]^T.$$

Os últimos rótulos n têm o valor de 0 para representar documentos não rotulados. O sistema de propagação também pode permitir a especificação de rótulos negativos, em vez de somente rótulos positivos, para representar exemplos negativos de relevância. O sistema de propagação representa distância entre documentos no espaço de características como $d: X \times X \rightarrow \mathbb{R}^+$, que atribui a cada par de pontos x_i e x_j uma distância $d(x_i, x_j)$, e representa uma função de classificação dos documentos como $f: X \rightarrow \mathbb{R}$, que atribui para cada ponto x_i uma marcação de classificação f_i . O problema da aprendizagem da função de classificação é aprender $f: X \rightarrow \mathbb{R}$ de um conjunto de consultas com as características $X = \{x_q\}$ e os rótulos $y = \{y_q\}$. O sistema de propagação representa o limite da propagação de relevância pela seguinte equação:

$$f^* = (1 - \alpha)(1 - \alpha S)^{-1} y \quad (4)$$

onde f^* representa o limite da relevância, y representa os rótulos iniciais, e α representa um fator de decomposição. Como é difícil em termos de computação calcular o inverso da matriz de similaridade normalizada S , o sistema de propagação aproxima f^* usando uma expansão da série de Taylor. O sistema de propagação pode representar a expansão da série de Taylor pela seguinte equação:

$$\begin{aligned} f^* &= (I - \alpha S)^{-1} y \\ &= (I + \alpha S + \alpha^2 S^2 + \dots) y \\ &= y + \alpha S y + \alpha S (\alpha S y) + \dots \end{aligned} \quad (5)$$

O sistema de propagação resolve iterativamente para f^* até que ele convirja em uma solução ou em um número fixo de iterações.

Após propagação das relevâncias, o sistema rotulado de propagação pode usar os conjuntos de dados de treinamento (consulta e vetores de características rotulados) para treinar uma função de classificação. Uma função de classificação pode ser implementada como uma máquina de vetor de suporte, um classificador de intensificação adaptável, um classificador de rede neural, e assim por diante. Uma máquina de vetor de suporte opera pelo encontro de uma hiper-superfície no espaço de entradas possíveis. A hiper-superfície tenta dividir os exemplos positivos a partir dos exemplos negativos maximizando a distância entre os exemplos negativos e positivos mais próximos para a hiper-superfície. Isto permite a classificação correta de dados que são similares, mas não idênticos aos dados de treinamento. Várias técnicas podem ser usadas para treinar uma máquina de vetor de suporte. Uma técnica usa um algoritmo de otimização mínima seqüencial que divide o problema de programação quadrática grande em uma série de problemas de programação quadrática pequenos que podem ser resolvidos analiticamente. (Veja Otimização Mínima seqüencial, em <http://reasearch.microsoft.com/~iplatt/smo.html>.)

A intensificação adaptável é um processo iterativo que roda testes múltiplos em uma coleção de dados de treinamento. A intensificação adaptável transforma um algoritmo

de aprendizagem fraco (um algoritmo que realiza em um nível somente um pouco melhor do que ocorre eventualmente) em um algoritmo de aprendizagem forte (um algoritmo que exibe uma taxa de erros baixa). O algoritmo de aprendizagem fraca é rodado em subconjuntos diferentes dos dados de treinamento. O algoritmo se concentra mais e mais naqueles exem-
 5 plos onde seus predecessores tenderam a mostrar erros. O algoritmo corrige os erros cometidos por aprendizes fracos anteriores. O algoritmo é adaptável porque ele se ajusta às taxas de erro de seus predecessores. A intensificação adaptável combina normas práticas aproximadas e moderadamente imprecisas para criar um algoritmo de alto desempenho. A intensificação adaptável combina os resultados de cada teste conduzido separadamente em
 10 um único classificador muito preciso.

Um modelo de rede neural tem três componentes principais: arquitetura, função de custo, e algoritmo de busca. A arquitetura define a forma funcional relacionando as entradas às saídas (em termos de topologia de rede, conectividade de unidade, e funções de ativação). A busca em espaço de peso para um conjunto de pesos que minimiza a função objeti-
 15 va é o processo de treinamento. Um modelo de rede neural pode usar uma rede de função de base radial ("RBF") e um declínio de gradiente padrão como sua técnica de busca.

A figura 1 é um diagrama que ilustra um gráfico de documentos retornados como o resultado da busca de uma consulta. Neste exemplo, o sub-gráfico 100 representa uma porção dos documentos retornados no resultado da busca. Os nós 101-112 representam 12
 20 documentos do resultado da busca. Os nós 101 e 106 representam documentos rotulados. O documento representado pelo nó 101 foi rotulado com a marcação de relevância de .75, e o documento representado pelo nó 106 foi rotulado com a marcação de relevância de .6. o sistema de propagação gerou as bordas entre os nós usando um algoritmo vizinho mais próximo. Neste exemplo, os nós 102, 103, e 104 são cada um dos vizinhos mais próximos
 25 de K para o nó 101, mas os nós 106-112 não são um dos vizinhos mais próximos de K. O sistema de propagação então calculou a similaridade entre nós conectados usando um algoritmo de marcação de similaridade. Por exemplo, o nó 101 é conectado ao nó 102 com uma borda com o peso de .8, que indica a similaridade entre os nós conectados.

A figura 2 é um diagrama de bloco que ilustra componentes do sistema de propagação em uma modalidade. O sistema de propagação 230 é conectado ao armazenamento de documentos 210 (por exemplo, sites da rede) por meio de links de comunicações 220 (por exemplo, a Internet). O sistema de propagação inclui um componente de coletar dados de treinamento 231, um armazenamento de dados de treinamento 232, e um índice de documentos 233. O índice de documentos contém um índice de documentos (por exemplo, páginas da rede) nos armazenamentos de documentos. O índice de documentos pode ser
 30 gerado por um arrastador da rede. O índice de documentos pode incluir um vetor de características para cada documento que é usado para treinar uma função de classificação. Os

vetores de características podem representar muitos tipos diferentes de características de documentos tais como frequência de documentos inversa, palavras-chaves, tamanho da fonte, e assim por diante. O componente de coletar dados de treinamento apresenta consultas a um motor de busca (não mostrado), e recebe documentos que correspondem às consultas. O motor de busca pode ser independente do sistema de propagação. Em tal caso, o sistema de propagação pode gerar vetores de características dinamicamente dos resultados das buscas. O componente de coletar dados de treinamento pode induzir um usuário a rotular a relevância de alguns dos documentos que correspondem às consultas. O componente de coletar dados de treinamento armazena as consultas, resultados das buscas (por exemplo, vetores de características), e rótulos no armazenamento de dados de treinamento. O sistema de propagação também inclui um componente de propagar relevância 235, um componente de construir gráfico 236, um componente de gerar pesos para gráfico 237, componente de normalizar pesos de gráfico 238 e um componente de propagar relevância baseada no componente gráfico 239. O componente de propagar relevância propaga a relevância dos documentos rotulados para os documentos não rotulados que estão armazenados no armazenamento de dados de treinamento. O componente de propagar relevância chama o componente de construir gráfico para construir um gráfico incluindo bordas representando os documentos de um resultado da busca. O componente de propagar relevância então chama o componente de gerar pesos para gráfico para gerar os pesos iniciais para as bordas do gráfico. O componente de propagar relevância chama o componente de normalizar pesos de gráfico para normalizar os pesos gerados. O componente de propagar relevância então chama o componente de propagar relevância com base em gráfico para executar a propagação real da relevância dos documentos rotulados em relação aos documentos não rotulados. O sistema de propagação também inclui um componente de criar função de classificação 241 e uma função de classificação 242. A função de criar classificação usa os dados de treinamento com a relevância propagada para criar uma função de classificação.

Os dispositivos de computação nos quais o sistema de propagação pode ser implementado podem incluir uma unidade de processamento central, memória, dispositivos de entrada (por exemplo, teclado e dispositivos de indicação), dispositivos de saída (por exemplo, dispositivos de exibição), e dispositivos de armazenamento (por exemplo, unidades de disco). Os dispositivos de memória e armazenamento são meios legíveis por computador que pode conter instruções que implementam o sistema de propagação. Além do mais, as estruturas de dados e estruturas de mensagens podem ser armazenadas ou transmitidas via um meio de transmissão de dados, tal como um sinal em um link de comunicações. Vários links de comunicações podem ser usados, tais como a Internet, uma rede de área local, uma rede remota, ou uma conexão de discagem de ponto a ponto.

O sistema de propagação pode fornecer serviços para vários sistemas ou dispositivos de computação incluindo computadores pessoais, computadores servidores, dispositivos de laptop ou de mão, sistemas de multiprocessadores, sistemas baseados em microprocessadores, eletrônica de consumidor programável, PCs de rede, minicomputadores, computadores de grande porte, ambientes de computação distribuídos que incluem quaisquer dos sistemas ou dispositivos acima, e similares.

O sistema de propagação pode ser descrito no contexto geral de instruções executáveis por computador, tais como módulos de programa, executadas por um ou mais computadores ou outros dispositivos. Geralmente, módulos de programas incluem rotinas, programas, objetos, componentes, estruturas de dados, e outros que realizam tarefas particulares ou implementam tipos de dados abstratos particulares. Tipicamente, a funcionalidade dos módulos dos programas pode ser combinada ou distribuída como desejado em várias modalidades.

A figura 3 é um diagrama de fluxo que ilustra o processamento do componente de criar função de classificação do sistema de propagação em uma modalidade. O componente de criar função de classificação coleta dados de treinamento, propaga a relevância de documentos rotulados para documentos não rotulados, e então treina uma função de classificação. No bloco 301, o componente coleta os dados de treinamento. No bloco 302, o componente entra com rótulos para um subconjunto dos dados de treinamento. No bloco 303, o componente chama o componente de propagar relevância para propagar a relevância dos documentos rotulados para os documentos não rotulados. No bloco 304, o componente treina a função de classificação usando relevâncias propagadas.

A figura 4 é um diagrama de fluxo que ilustra o processamento do componente de propagar relevância do sistema de propagação em uma modalidade. O componente é suprido com dados de treinamento e propaga a relevância dos documentos rotulados para os documentos não rotulados. No bloco 401, o componente chama o componente de construir gráfico para construir o gráfico inicial que inclui bordas. No bloco 402, o componente chama o componente de gerar pesos para gráfico para gerar pesos indicando a similaridade entre documentos representados pelos nós conectados. No bloco 403, o componente chama o componente de normalizar pesos de gráfico para normalizar os pesos do gráfico. No bloco 404, o componente chama o componente de propagar relevância com base em gráfico para realizar a propagação da relevância. O componente depois retorna.

A figura 5 é um diagrama de fluxo que ilustra o processamento do componente de construir gráfico do sistema de propagação em uma modalidade. O componente cria uma matriz quadrada com cada linha e coluna representando um documento. O componente então identifica e adiciona uma conexão entre cada nó e seus vizinhos mais próximos de K (por exemplo, $K=10$). No bloco 501, o componente seleciona o próximo documento i . No bloco de

decisão 502, se todos os documentos i já tiverem sido selecionados, então o componente retorna, caso contrário o componente continua no bloco 503. No bloco 503, o componente seleciona o próximo documento j . No bloco de decisão 504, se todos os documentos j para o documento selecionado i já tiverem sido selecionados, então o componente continua no bloco 506, caso contrário o componente continua no bloco 505. No bloco 505, o componente calcula a distância entre o documento selecionado i e o documento selecionado j e então a reverte para o bloco 503 para selecionar o próximo documento j . No bloco 505, o componente seleciona os 10 documentos j com a menor distância para um documento i (isto é, os vizinhos mais próximos) e então reverte para o bloco 501 para selecionar o próximo documento i .

A figura 6 é um diagrama de fluxo que ilustra o processamento do componente de gerar pesos para gráfico do sistema de propagação em uma modalidade. O componente calcula a similaridade entre documentos conectados baseado em uma métrica de Manhattan. No bloco 601, o componente seleciona o próximo documento i . No bloco de decisão 602, se todos os documentos i já tiverem sido selecionados, então o componente retorna, caso contrário o componente continua no bloco 603. No bloco 603, o componente inicializa a similaridade do documento para si mesmo para 0. No bloco 604, o componente seleciona o documento seguinte j mais próximo (isto é, um documento conectado) para o documento selecionado i . No bloco de decisão 605, se todos os documentos j mais próximos ao documento selecionado i já tiverem sido selecionados, então o componente reverte para o bloco 601 para selecionar o próximo documento i , caso contrário o componente continua no bloco 606. No bloco 606, o componente inicializa a similaridade entre o documento selecionado i e o documento selecionado j para 1. Nos blocos 607-609, o componente reverte calculando a métrica de distância. No bloco 607, o componente seleciona a próxima dimensão i do vetor de características. No bloco de decisão 606, se todas as dimensões já tiverem sido selecionadas, então o componente reverte para o bloco 604 para selecionar o próximo documento j mais próximo, caso contrário o componente continua no bloco 609. No bloco 609, o componente estabelece a similaridade entre o documento selecionado i e o documento selecionado j para a sua similaridade atual multiplicada por uma função da diferença entre a característica selecionada i do documento selecionado i e o documento selecionado j de acordo com a Equação 2. O componente então reverte para o bloco 607 para selecionar a próxima dimensão.

A figura 7 é um diagrama de fluxo que ilustra o processamento do componente de normalizar pesos de gráfico do sistema de propagação em uma modalidade. O componente normaliza os pesos da matriz de similaridade. No bloco 701, o componente seleciona a próxima linha i da matriz de similaridade. No bloco de decisão 702, se todas as linhas já tiverem sido selecionadas, então o componente continua no bloco 706, caso contrário o componente

continua no bloco 703. Nos blocos 703-705, o componente calcula o valor para a matriz diagonal D para a linha selecionada. No bloco 703 o componente seleciona a próxima coluna j da matriz de similaridade. No bloco de decisão 704, se todas as colunas já tiverem sido selecionadas, então o componente reverte para o bloco 701 para selecionar a próxima linha, caso contrário o componente continua no bloco 705. No bloco 705, o componente adiciona os pesos da linha selecionada i a coluna selecionada j para o elemento diagonal para a linha selecionada i . O componente então reverte para o bloco 703 para selecionar a próxima coluna j para a linha selecionada i . No bloco 706, o componente normaliza a matriz de similaridade de acordo com a Equação 3.

A figura 8 é um diagrama de fluxo que ilustra o processamento do componente de propagar relevância baseada em gráfico do sistema de propagação em uma modalidade. O componente calcula iterativamente a expansão da série de Taylor da Equação 5 até que ela convirja em uma solução. No bloco 801, o componente inicializa o índice i para zero. No bloco 802, o componente inicializa o vetor de solução para 0. Nos blocos 803-805, o componente reverte até que ele convirja para uma solução. No bloco 803, o componente calcula o valor para a próxima iteração baseado no valor da iteração anterior mais o próximo fator da expansão da série de Taylor. No bloco de decisão 804, se os valores convergem para uma solução, então o componente retorna, caso contrário o componente continua no bloco 805. No bloco 805, o componente incrementa o índice para a próxima iteração e reverte para o bloco 803 para realizar a próxima iteração.

Embora o assunto tenha sido descrito em linguagem específica para características estruturais e/ou atos metodológicos, deve ser entendido que o assunto definido nas reivindicações anexadas não está necessariamente limitado às características ou atos específicos descritos acima. Antes, as características e atos específicos descritos acima são revelados como formas de exemplo de implementação das reivindicações. O sistema de propagação pode ser usado para aumentar resultados de busca. Por exemplo, um motor de busca pode gerar um resultado de busca baseado em certo corpo de documentos. A relevância dos documentos do resultado da busca pode então ser propagada para documentos de um corpo diferente usando o sistema de propagação. Os documentos do corpo diferente com a relevância mais alta podem então ser adicionados ao resultado da busca. O sistema de propagação pode ser usado para propagar a relevância de documentos rotulados com sua relevância para uma consulta única a documentos não rotulados (propagação intraconsulta) ou de documentos rotulados com sua relevância a consultas múltiplas para documentos não rotulados (propagação inter-consulta). O componente de propagação treina o componente de treinamento separadamente para cada consulta com propagação intra-consulta e simultaneamente para consultas múltiplas com propagação inter-consultas. Consequentemente, a invenção não é limitada exceto pelas reivindicações anexadas.

REIVINDICAÇÕES

1. Sistema para treinar um componente de classificação de documentos, **CARACTERIZADO** por compreender:

5 Um armazenamento de dados de treinamento (232) que contém dados de treinamento incluindo representações de documentos e, para cada consulta de uma pluralidade de consultas, uma rotulação de alguns documentos com relevância dos documentos para a consulta;

10 Um componente de propagar relevância (235) que propaga a relevância dos documentos rotulados para os documentos não rotulados baseados na similaridade entre documentos; e

Um componente de treinamento (241) que treina um componente de classificação de documentos para classificar a relevância de documentos para consultas baseado na relevância propagada dos documentos dos dados de treinamento.

15 2. Sistema de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de classificação de documentos implementa um algoritmo de classificação selecionado de um grupo consistindo em um algoritmo de rede neural, um algoritmo de intensificação adaptável, e um algoritmo de máquina de vetor de suporte.

20 3. Sistema de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de classificação de documentos implementa um algoritmo baseado em regressão.

4. Sistema de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de propagar relevância propaga a relevância separadamente para cada consulta e o componente de treinamento treina o componente de classificação de documentos usando as relevâncias propagadas separadamente.

25 5. Sistema de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de propagar relevância propaga a relevância simultaneamente para consultas múltiplas e o componente de treinamento treina o componente de classificação de documentos usando as relevâncias propagadas simultaneamente.

30 6. Sistema de acordo com a reivindicação 1 **CARACTERIZADO** por incluir um componente gráfico que cria um gráfico com os documentos representados como nós conectados por bordas representando a similaridade entre documentos.

7. Sistema de acordo com a reivindicação 6, **CARACTERIZADO** pelo fato de que o componente gráfico inclui:

35 Um componente de construir gráfico que constrói um gráfico no qual nós representando documentos similares são conectados por meio de bordas; e

Um componente de gerar pesos que gera pesos para as bordas baseado na similaridade dos documentos representados pelos nós conectados.

8. Sistema de acordo com a reivindicação 7, **CARACTERIZADO** pelo fato de que o componente de construir gráfico estabelece bordas entre nós usando um algoritmo vizinho mais próximo.

5 9. Sistema de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de propagar relevância propaga a relevância usando um algoritmo baseado em classificação de tubulação.

10. Meio legível por computador contendo instruções para controlar um sistema de computador para treinar um componente de classificação de documentos, por um método **CARACTERIZADO** por compreender:

10 Fornecer (232) representações de documentos junto com uma rotulação de alguns dos documentos que indica relevância de um documento para uma consulta,

Criar (236) um gráfico com os documentos representados como nós sendo conectados por bordas representando a similaridade entre documentos representados pelos nós conectados;

15 Propagar (239) a relevância dos documentos rotulados para os documentos não rotulados baseado na similaridade entre documentos como é indicado pelo gráfico criado e baseado em um algoritmo baseado em classificação múltipla; e

Treinar (241) um componente de classificação de documentos para classificar a relevância de documentos para consultas baseada na relevância propagada dos documentos.

20 11. Meio legível por computador de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que o componente de classificação de documentos implementa um algoritmo de classificação selecionado de um grupo consistindo em um algoritmo de rede Bayes, um algoritmo de intensificação adaptável, e um algoritmo de máquina de vetor de suporte.

25 12. Meio legível por computador de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que o componente de classificação de documentos implementa um algoritmo de classificação baseado em regressão.

30 13. Meio legível por computador de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que a propagação da relevância propaga a relevância separadamente para cada consulta e o treinamento do componente de classificação de documentos treina o uso da relevância propagada separadamente.

14. Meio legível por computador de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que o componente de propagar relevância propaga a relevância [propagação inter-consulta].

35 15. Meio legível por computador de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que a criação de um gráfico inclui:

Construir um gráfico no qual nós representando documentos similares são conecta-

dos por meio de bordas; e

Gerar pesos para as bordas com base na similaridade dos documentos representados pelos nós conectados.

16. Sistema para treinar um componente de classificação de documentos,

5 **CARACTERIZADO** por compreender:

Um componente (231) que fornece representações de documentos junto com uma rotulação de alguns dos documentos que indica a relevância dos documentos para consultas;

10 Um componente (236) que cria um gráfico com os documentos representados como nós sendo conectados por bordas representando similaridade entre documentos representados pelos nós conectados;

Um componente (239) que propaga a relevância dos documentos rotulados para os documentos não rotulados baseado na similaridade entre documentos como indicado pelo gráfico criado; e

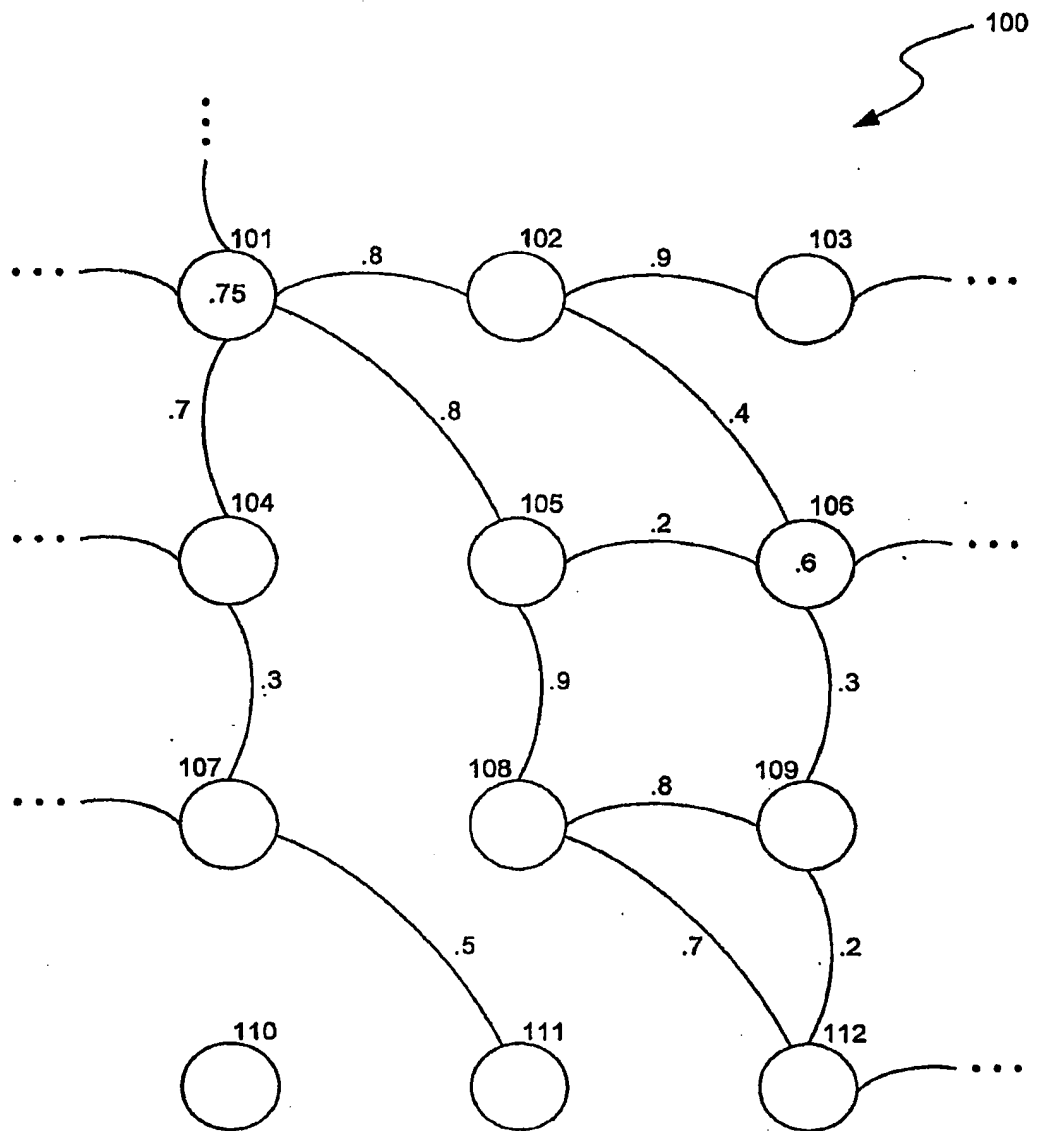
15 Um componente que gera um componente de classificação de documentos para classificar a relevância de documentos para consultas baseado na relevância propagada dos documentos.

20 17. Sistema de acordo com a reivindicação 16, **CARACTERIZADO** pelo fato de que o componente que propaga relevância propaga relevância baseado em um algoritmo baseado em classificação múltipla.

18. Sistema de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que o componente que propaga relevância propaga relevância simultaneamente para consultas múltiplas e o componente que gera o componente de classificação de documentos gera o componente usando a relevância propagada simultaneamente.

25 19. Sistema de acordo com a reivindicação 16, **CARACTERIZADO** pelo fato de que o componente que cria um gráfico constrói um gráfico, gera pesos para as bordas baseado na similaridade dos documentos representados pelos nós conectados.

30 20. Sistema de acordo com a reivindicação 16, **CARACTERIZADO** pelo fato de que o componente de classificação de documentos implementa um algoritmo de classificação baseado em regressão.

**FIG. 1**

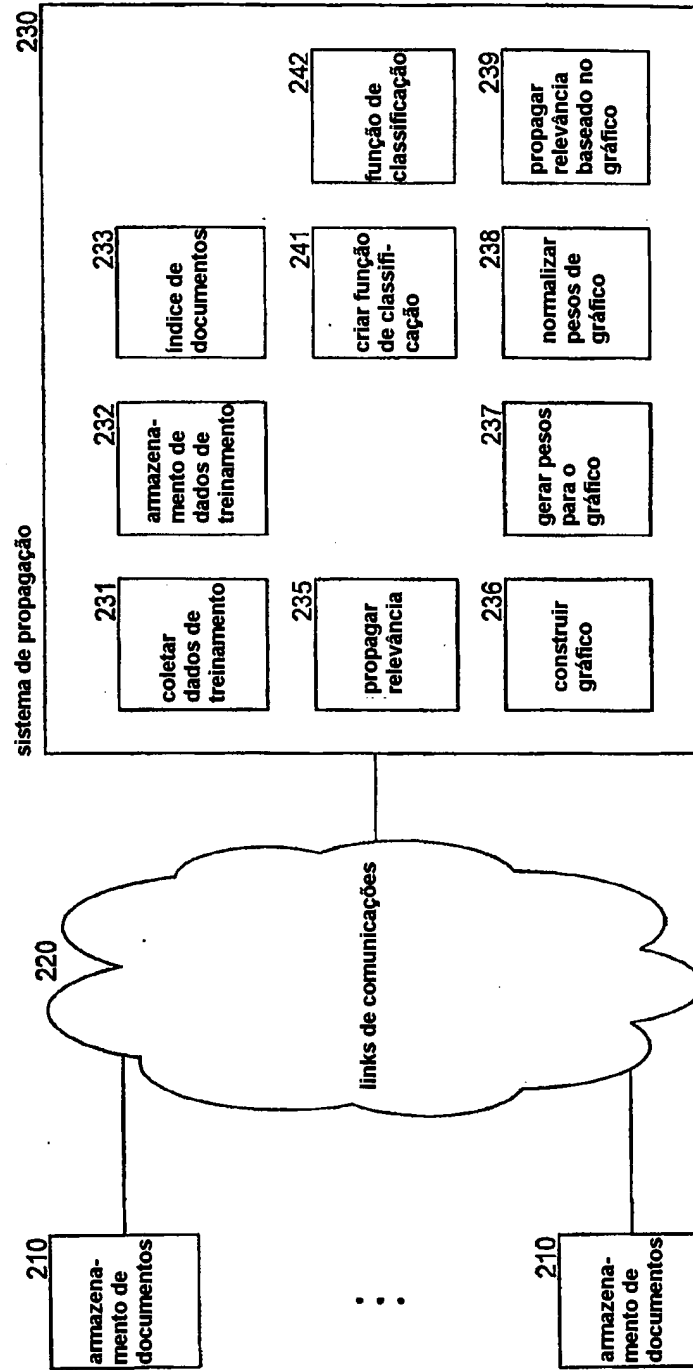
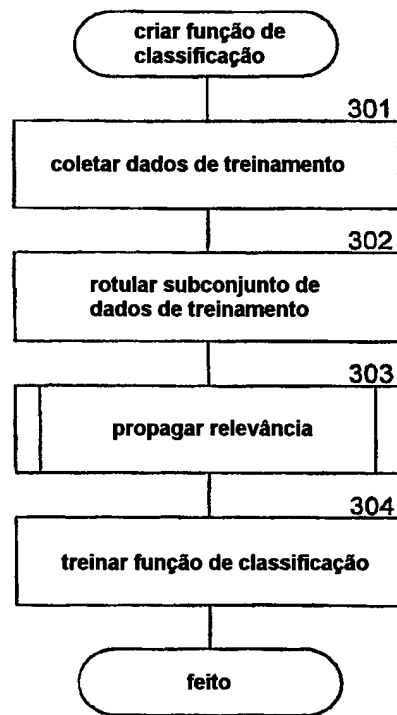
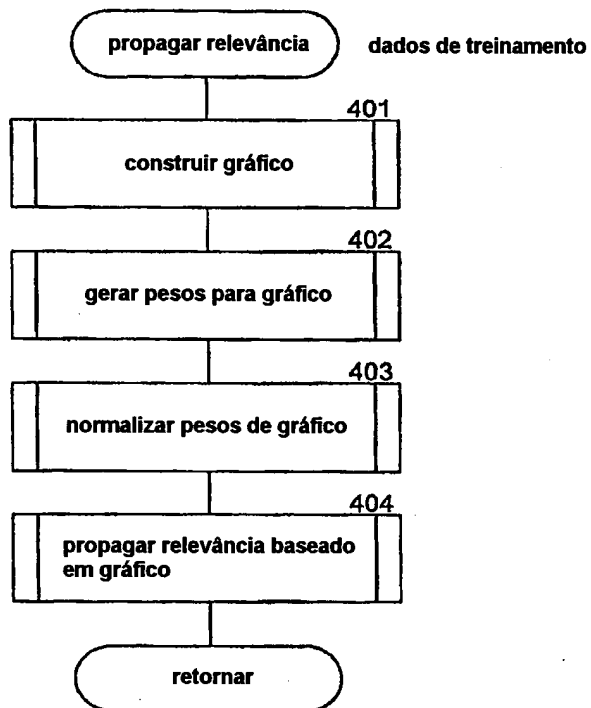
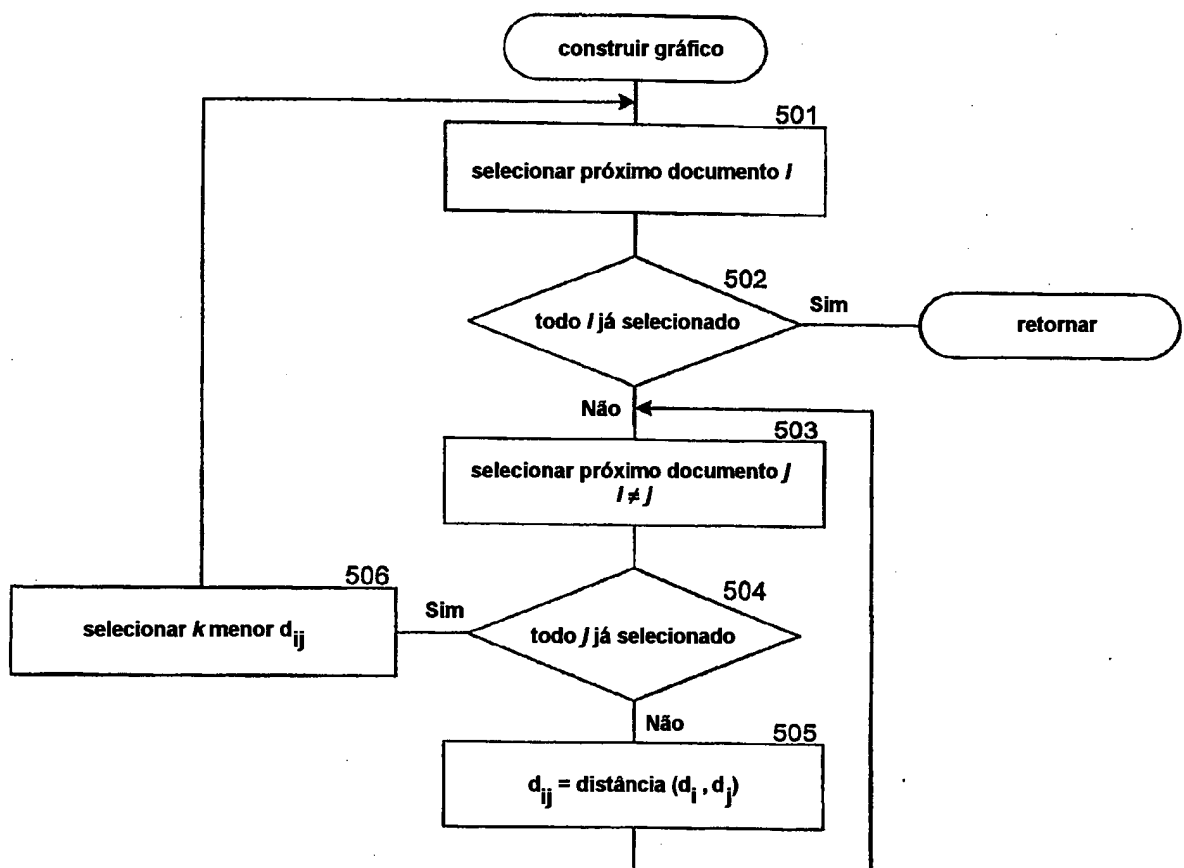


FIG. 2

**FIG. 3**

**FIG. 4**

**FIG. 5**

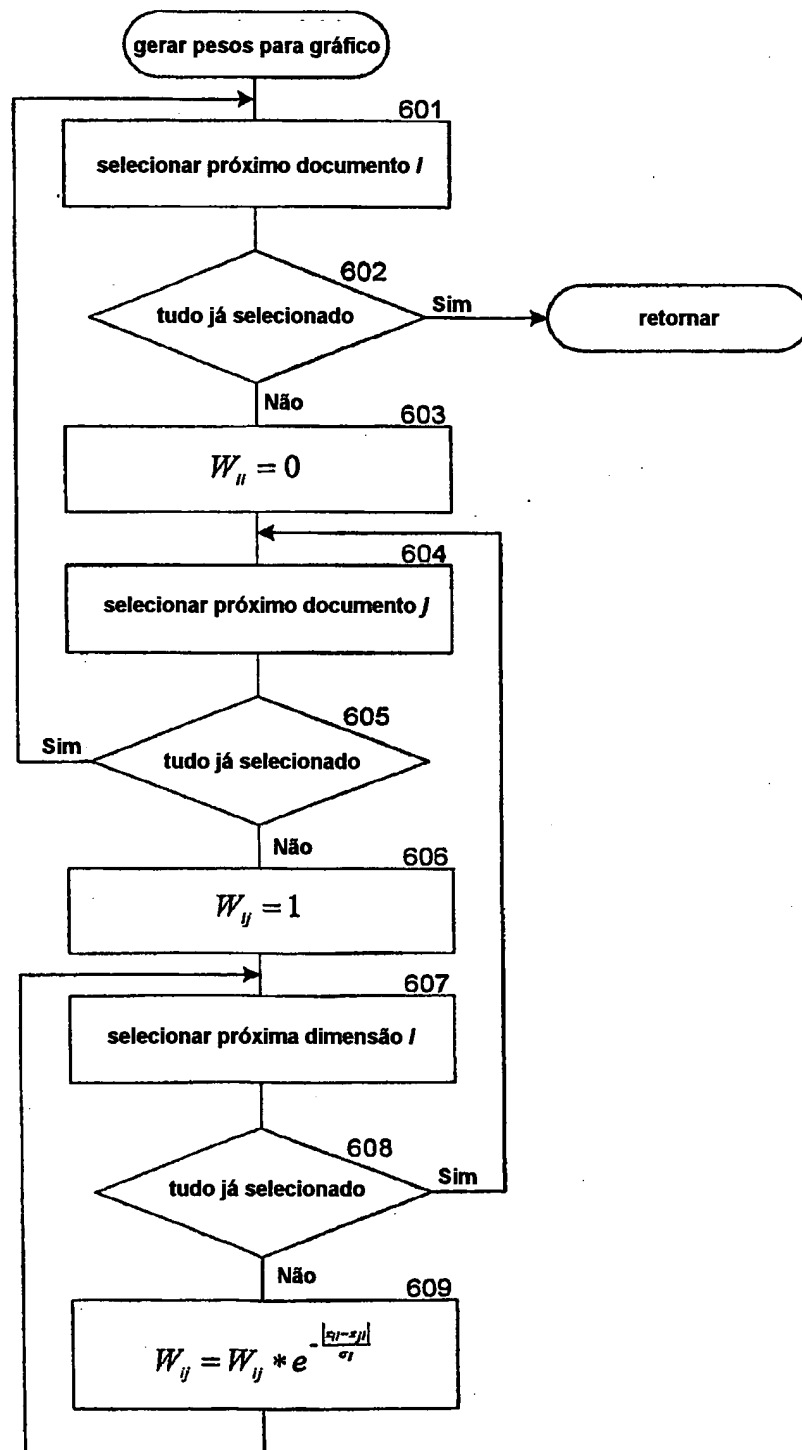
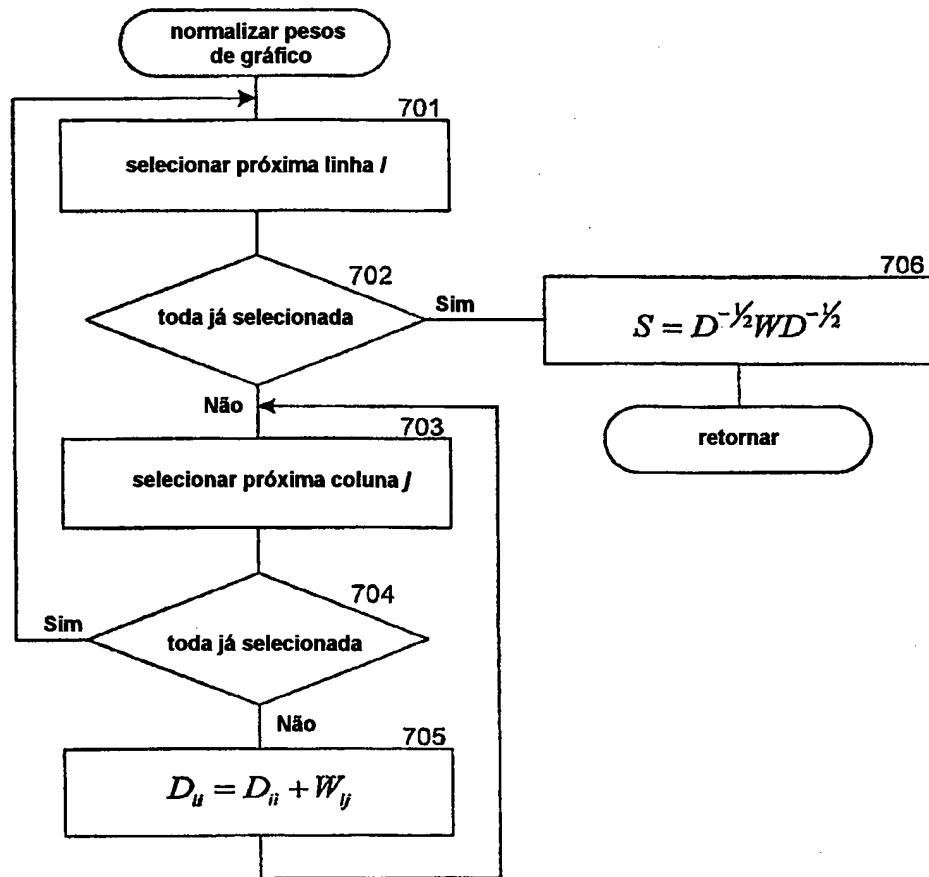
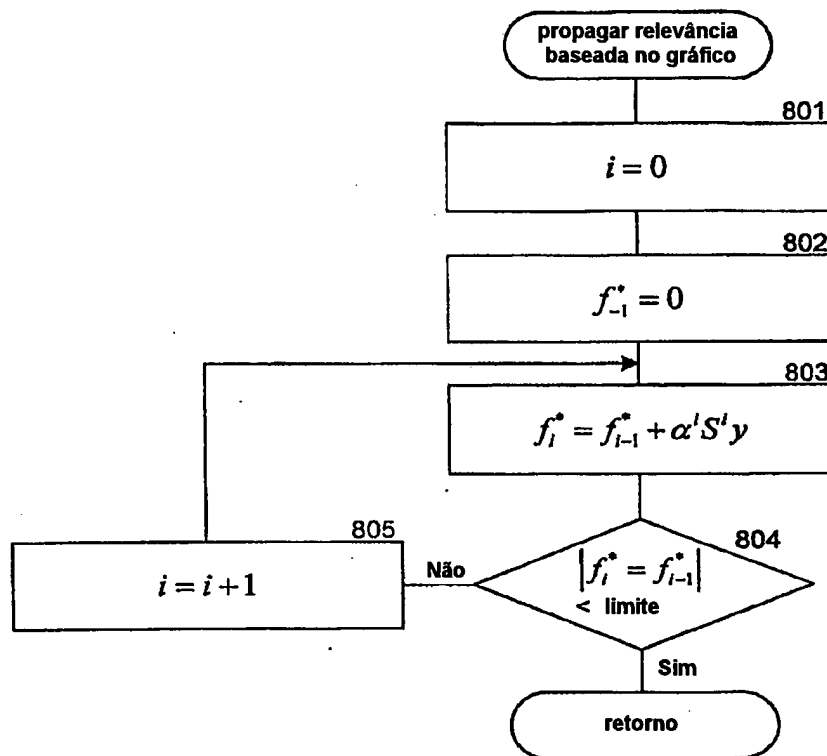


FIG. 6

**FIG. 7**

**FIG. 8**

RESUMO**"TREINAMENTO DE UMA FUNÇÃO DE CLASSIFICAÇÃO USANDO RELEVÂNCIA PROPAGADA DE DOCUMENTOS"**

Um método e sistema para propagar a relevância de documentos rotulados para uma consulta para documentos não rotulados são fornecidos. O sistema de propagação fornece dados de treinamento que inclui consultas, documentos rotulados com sua relevância para as consultas, e documentos não rotulados. O sistema de propagação então calcula a similaridade entre pares de documentos nos dados de treinamento. O sistema de propagação então propaga a relevância dos documentos rotulados para documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos converjam para uma solução. Os dados de treinamento com as relevâncias propagadas podem então ser usados para treinar uma função de classificação.