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

(21) **PI0708074-3 A2**

* B R P I 0 7 0 8 0 7 4 A 2 *

(22) Data de Depósito: 27/02/2007
(43) Data da Publicação: 17/05/2011
(RPI 2106)

(51) Int.Cl.:
G06F 17/30

(54) Título: **PROPAGAÇÃO DE RELEVÂNCIA DE DOCUMENTOS ROTULADOS PARA DOCUMENTOS NÃO ROTULADOS**

(30) Prioridade Unionista: 27/02/2006 US 11/364.807

(73) Titular(es): Microsoft Corporation

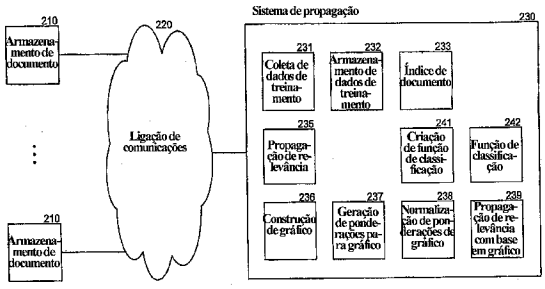
(72) Inventor(es): Jue Wang, Mingjing Li, Wei-Ying Ma, Zhiwei Li

(74) Procurador(es): Alexandre Ferreira

(86) Pedido Internacional: PCT US2007005149 de 27/02/2007

(87) Publicação Internacional: WO 2007/100848de 07/09/2007

(57) Resumo: PROPAGAÇÃO DE RELEVÂNCIA DE DOCUMENTOS ROTULADOS PARA DOCUMENTOS NÃO ROTULADOS. São fornecidos um método e sistema para propagar a relevância de documentos rotulados em relação a uma consulta até documentos não rotulados. O sistema de propagação fornece dados de treinamento que incluem consultas, documentos rotulados com suas relevâncias em relação às consultas, e documentos não rotulados. Então, o sistema de propagação calcula a similaridade entre pares de documentos nos dados de treinamento. Então, o sistema de propagação propaga a relevância dos documentos rotulados em documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos convirjam em uma solução. Então, os dados de treinamento com as relevâncias propagadas podem ser usados para treinar uma função de classificação.



"PROPAGAÇÃO DE RELEVÂNCIA DE DOCUMENTOS ROTULADOS PARA DOCUMENTOS NÃO ROTULADOS"

ANTECEDENTES DA INVENÇÃO

Muitos serviços de motor de busca, tais como Google e Overture, fornecem busca de informação que é acessível por meio da Internet. Estes serviços de motor de busca permitem que usuários busquem páginas de exibição, tais como páginas da Internet, que podem ser de interesse dos usuários. Depois que um usuário submete uma solicitação de busca (isto é, uma consulta) que inclui termos de busca, o serviço de motor de busca identifica páginas da Internet que podem estar relacionadas àqueles termos da busca. Para identificar rapidamente páginas da Internet relacionadas, os serviços de motor de busca podem manter um mapeamento de palavras-chaves para as páginas da Internet. Este mapeamento pode ser gerado pelo "esquadrinhamento" da Internet (isto é, a rede mundial de computadores) para identificar as palavras-chaves de cada página da Internet. Para esquadrinhar a Internet, um serviço de motor de busca pode usar uma lista de páginas raízes da Internet para identificar todas as páginas da Internet que são acessíveis por meio de daquelas páginas raízes da Internet. As palavras-chaves de qualquer página da Internet em particular podem ser identificadas usando várias técnicas de recuperação de informação bem conhecidas, tais como identificação das palavras de um cabeçalho, as palavras supridas nos metadados da página da Internet, as palavras que estão destacadas, e assim por diante. O serviço de motor de busca identifica páginas da Internet que podem ser relacionadas à solicitação de busca com base em quão bem as palavras-chaves de uma página da Internet casam com as palavras da consulta. Então, o serviço de motor de busca exibe ao usuário ligações para as páginas da Internet identificadas em uma ordem que é baseada em uma classificação que pode ser determinada por suas relevâncias em relação à consulta, à popularidade, à importância e/ou a alguma outra medida.

Estas técnicas bem conhecidas para classificar páginas da Internet são PageRank, HITS ("Busca de Tópico Induzida por Hiperligação"), e DirectHIT. PageRank é baseada no princípio de que páginas da Internet terão ligações para (isto é, "ligações de saída") importantes páginas da Internet. Assim, a importância da página da Internet é baseada no número e na importância de outras páginas da Internet que se ligam àquela página da Internet (isto é, "ligações de entrada"). De uma forma simples, as ligações entre as páginas da Internet podem ser representadas pela matriz de adjacência A, em que A_{ij} representa o número de ligações de saída da página da Internet i até a página da Internet j. A contagem de importância w_j para a página da Internet j pode ser representada pela seguinte equação:

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

Esta equação pode ser resolvida por cálculos iterativos com base na seguinte equação:

$A^T w = w$

em que w é o vetor de contagens de importância para as páginas da Internet e é o principal autovetor de A^T.

Esta técnica HITS é adicionalmente baseada no princípio de que uma própria página da Internet que tem muitas ligações a outras importantes páginas da Internet pode ser importante. Assim, HITS divide "importância" das páginas da Internet em dois atributos relacionados, "concentrador" e "autoridade". "Concentrador" é medido pela contagem de "autoridade" das páginas da Internet em que uma página da Internet se liga, e "autoridade" é medida pela contagem do "concentrador" das páginas da Internet que se ligam à página da Internet. Ao contrário da PageRank, que calcula a importância das páginas da Internet independentemente da consulta, HITS calcula a importância com base nas páginas da Internet do resultado e das páginas da Internet que são relacionadas às páginas da Internet do resultado seguindo as ligações de entrada e de saída. HITS submete uma consulta a um serviço de motor de busca e usa as páginas da Internet do resultado como o conjunto inicial de páginas da Internet. HITS adiciona no conjunto aquelas páginas da Internet que são destinos das ligações de entrada e aquelas páginas da Internet que são as fontes das ligações de saída das páginas da Internet do resultado. Então, HITS calcula a classificação da autoridade e do concentrador de cada página da Internet usando um algoritmo iterativo. As contagens da autoridade e do concentrador podem ser representadas pelas seguintes equações:

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

em que $a(p)$ representa a contagem da autoridade para a página da Internet p e $h(p)$ representa a contagem do concentrador para a página da Internet p. HITS usa uma matriz de adjacência A para representar as ligações. A matriz de adjacência é representada pela seguinte equação:

$$B_{ij} = \begin{cases} 1 \text{ se página } i \text{ tiver uma ligação até a página } j, \\ 0 \text{ caso contrário} \end{cases}$$

Os vetores a e h correspondem às contagens da autoridade e do concentrador, respectivamente, de todas as páginas da Internet no conjunto e podem ser representados pelas seguintes equações:

$A = A^T h$ e $h = Aa$

Assim, a e h são autovetores das matrizes A^TA e AA^T. HITS também pode ser modificado para o fator na popularidade de uma página da Internet medida pelo número de visitas. Com base na análise dos dados através de cliques, b_{ij} da matriz de adjacência pode aumentar toda vez que um usuário navega da página da Internet i até a página da Internet j.

DirectHIT classifica páginas da Internet com base no histórico passado do usuário

com resultados de consultas similares. Por exemplo, se usuários que submetem consultas similares selecionam primeiro, tipicamente, a terceira página da Internet do resultado, então, o histórico deste usuário será uma indicação de que a terceira página da Internet deve ter classificação mais alta. Como um outro exemplo, se usuários que submetem consultas similares gastam, tipicamente, a maior parte do tempo visualizando a quarta página da Internet do resultado, então, o histórico deste usuário será uma indicação de que a quarta página da Internet deve ter classificação mais alta. DirectHIT deriva os históricos do usuário da análise dos dados através do clique.

Algumas técnicas de classificação usam algoritmos de aprendizado de máquina para aprender uma função de classificação dos dados de treinamento que incluem consultas, vetores de recurso que representam páginas e, para cada consulta, uma classificação para cada página. Uma função de classificação serve como um mapeamento dos recursos de uma página para sua classificação para uma dada consulta. O aprendizado de uma função de classificação foi considerado, em parte, como um problema de regressão para aprender o mapeamento de um vetor de recurso até um elemento de um conjunto ordenado de classificações numéricas. Algumas técnicas com base em regressão tentam fornecer uma contagem de relevância absoluta que pode ser usada para classificar páginas. Entretanto, uma função de classificação não precisa fornecer uma contagem de relevância absoluta, mas, em vez disto, precisa fornecer somente uma classificação relativa das páginas. Assim, estas técnicas com base em regressão resolvem um problema que é mais difícil do que o necessário.

Algoritmos de aprendizagem de máquina para uma função de classificação usam consultas, vetores de recurso e classificações de relevância rotuladas pelo usuário como dados de treinamento. Para gerar os dados de treinamento, consultas podem ser submetidas a um motor de busca que gera as páginas do resultado da busca. Então, os algoritmos geram os vetores de recurso para as páginas e inserem, a partir de um usuário, as contagens de relevância para cada página. Uma dificuldade com uma abordagem como esta é que um motor de busca pode retornar centenas de páginas como seu resultado de busca. Pode ser bastante oneroso ter um rótulo de usuário em todas as páginas de um resultado de busca. Além do mais, pode ser difícil para um usuário avaliar precisamente a relevância de um grande número de páginas como este. Embora um usuário possa rotular somente uma pequena parte das páginas, o aprendizado com base em uma pequena parte como esta pode não fornecer uma função de classificação precisa.

SUMÁRIO DA INVENÇÃO

São fornecidos um método e sistema para propagar a relevância dos documentos rotulados a uma consulta para a relevância dos documentos não rotulados. Este sistema de propagação fornece dados de treinamento que incluem consultas, documentos rotulados

com suas relevâncias em relação às consultas e documentos não rotulados. Então, o sistema de propagação calcula a similaridade entre pares de documentos nos dados de treinamento. Então, o sistema de propagação propaga a relevância dos documentos rotulados a documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos convirjam em uma solução. Então, os dados de treinamento com as relevâncias propagadas podem ser usados para treinar uma função de classificação.

Este Sumário é fornecido para introduzir uma seleção de conceitos de uma forma simplificada que é adicionalmente descrita a seguir na Descrição Detalhada. Não pretende-se que este Sumário identifique recursos chaves ou recursos essenciais do assunto em questão reivindicado, nem pretende-se que seja usado como um auxílio na determinação do escopo do assunto em questão reivindicado.

DESCRIÇÃO RESUMIDA DOS DESENHOS

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

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

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

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

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

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

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

A figura 8 é um fluxograma que ilustra o processamento do componente de propagar relevância com base em gráfico do sistema de propagação em uma modalidade.

DESCRIÇÃO DETALHADA

São fornecidos um método e sistema para propagar regressão estatística de documentos rotulados a uma consulta para documentos não rotulados. Em uma modalidade, o sistema de propagação fornece dados de treinamento que incluem consultas, documentos (representados por vetores de recurso) rotulados com suas relevâncias em relação às consultas, e documentos não rotulados. Por exemplo, o sistema de propagação pode submeter uma consulta a um motor de busca e usar o resultado da busca como os documentos (por exemplo, páginas da Internet). Então, o sistema de propagação pode solicitar que um usuário rotule parte dos documentos do resultado da busca com base em suas relevâncias em

relação à consulta. Então, o sistema de propagação calcula a similaridade entre pares de documentos nos dados de treinamento. Por exemplo, o sistema de propagação pode representar cada documento por um vetor de recurso e pode calcular a similaridade entre documentos com base na distância Euclidiana no espaço do recurso ou com base em uma métrica de similaridade co-seno. Então, o sistema de propagação 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. Então, os dados de treinamento com as relevâncias propagadas podem ser usados para treinar uma função de classificação. Desta maneira, o sistema de propagação pode aumentar automaticamente os dados de treinamento com dados de treinamento adicionais com base 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 seta representando a similaridade entre os documentos representados pelos nós conectados. O sistema de propagação pode representar o gráfico como uma matriz quadrática com uma linha e coluna para cada documento na qual cada valor não zero indica uma seta entre o nó da linha e o nó da coluna. O sistema de propagação pode definir setas para o gráfico usando várias técnicas. Por exemplo, o sistema de propagação pode considerar que o gráfico está completamente conectado, em cujo caso cada nó tem uma seta para cada outro nó. Como um outro exemplo, o sistema de propagação pode considerar que os nós estão conectados por meio de uma árvore de abrangência mínima. Em uma modalidade, o sistema de propagação considera que os nós estão conectados usando um algoritmo do vizinho k mais próximo. Em particular, o sistema de propagação identifica os k vizinhos mais próximos para cada nó e adiciona uma seta de cada nó até cada um dos seus k vizinhos mais próximos. Então, o sistema de propagação calcula ponderações para as setas com base na similaridade entre os documentos representados pelas setas conectadas. O sistema de propagação pode usar várias técnicas para determinar a similaridade entre os documentos. Em uma modalidade, o sistema de propagação usa uma métrica de distância Euclidiana com base na representação do vetor de recurso dos documentos em um espaço do recurso. O sistema de propagação armazena a similaridade como os valores da matriz quadrática resultante em uma matriz de similaridade ou de afinidade. O sistema de propagação também pode normalizar a matriz de similaridade. O sistema de propagação também pode ajustar os valores diagonais em 0 ou impedir auto-reforço durante a propagação da relevância.

Depois de gerar a matriz de similaridade, o sistema de propagação propaga a relevância dos documentos rotulados até os documentos não rotulados usando um algoritmo de propagação com base em classificação de cópias. Um algoritmo com base em classificação de cópias é descrito em He, J., Li, M., Zhang, H.J., et al., "Manifold-Ranking Based Image

Retrieval", Proc. of the 12th Annual ACM International Conf. On Multimedia, 2004. Inicialmente, o sistema de propagação ajusta a regressão estatística dos documentos rotulados da contagem de relevância fornecida pelo usuário e a relevância dos documentos não rotulados em 0. Então, o sistema de propagação difunde a relevância dos documentos rotulados a seus documentos não rotulados conectados fatorando na similaridade como indicado pela matriz de similaridade. O sistema de propagação difunde iterativamente a contagem de relevância até que as contagens de relevância convirjam em uma solução. As contagens de relevância resultantes dos documentos não rotulados estarão em proporção com a probabilidade de que eles sejam relevantes à mesma consulta que os documentos rotulados. Assim, um documento não rotulado que é muito similar a muitos documentos rotulados com altas contagens de relevância terão uma alta contagem de relevância. Inversamente, um documento não rotulado que não é muito similar a nenhum documento rotulado terá uma baixa contagem de relevância.

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

(copiar fórmula pg 7)(1)

em que x_{il} e x_{jl} representam a $l^{ésima}$ dimensão de x_i e de x_j, respectivamente, t representa a dimensionalidade do espaço do recurso, e ρ_l representa um parâmetro positivo que reflete as ponderações das diferentes dimensões no cálculo da similaridade. Assim, o sistema de propagação representa a ponderação das setas pela seguinte equação:

(copiar fórmula pg 7)(2)

em que W_{ij} representa a similaridade entre os documentos i e j. O sistema de propagação pode omitir o coeficiente constante $1/2\rho_l$ desde que seu efeito na matriz de similaridade W seja neutralizado pela normalização da matriz. O sistema de propagação normaliza a matriz de similaridade como representado pela equação:

$S = D^{-1/2}WD^{-1/2}$(3)

em que S representa a matriz de similaridade normalizada e D representa uma matriz diagonal em que (i,i) é igual à soma da i-*ésima* linha da matriz de similaridade W. A normalização normaliza as similaridades para ser relativa à similaridade dos documentos conectados.

O sistema de propagação pode representar cada documento como um vetor de recurso χ da dimensão t que forma um ponto no espaço Euclidiano. Para uma consulta, o sistema de propagação recebe o conjunto de resultado dos documentos (copiar fórmula pg 8). Os primeiros m pontos (no espaço de recurso) representam documentos rotulados pelo usuário, e os últimos n pontos (no espaço de recurso) representam documentos não rotulados. O sistema de propagação também recebe um vetor de rótulo correspondente (copiar fórmula pg 8). Os últimos n rótulos têm o valor de 0 para representar documentos não rotu-

lados. O sistema de propagação também pode permitir a especificação dos 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 recurso como d: $\chi \times \chi \rightarrow \square$, que atribui a cada par de pontos χi e χj, uma distância $d(xi, xj)$, e representa uma função de classificação dos documentos como f: $\chi \rightarrow \square$, que atribui a cada ponto χi, uma contagem de classificação fi. O problema do aprendizado da função de classificação é aprender f: $\chi \rightarrow \square$ de um conjunto de consultas com os recursos $X = \{\chi_q\}$ e os rótulos $Y = \{\gamma_q\}$. O sistema de propagação representa o limite da propagação da relevância pela seguinte equação:

$$f^\circ = (1-\alpha)(I-\alpha S)^{-1}y(4)$$

em que f° representa o limite da relevância, y representa os rótulos iniciais e α representa um fator de decadência. Em virtude de ser computacionalmente difícil de calcular o inverso da matriz de similaridade normalizada S, o sistema de propagação aproxima f° usando uma expansão da série Taylor. O sistema de propagação pode representar a expansão da série Taylor pela seguinte equação:

$$\begin{aligned} f^\circ &= (I - \alpha S)^{-1}y \\ &= (I + \alpha S + \alpha^2 S^2 + \ldots)y(5) \\ &= y + \alpha Sy + \alpha S(\alpha Sy) + \ldots \end{aligned}$$

O sistema de propagação resolve iterativamente f° até que ela convirja em uma solução ou por um número fixo de iterações.

Uma vez que as relevâncias são propagadas, o sistema de propagação rotulada pode usar os conjuntos de dados de treinamento (vetores de recurso de consulta e rotulados) para treinar uma função de classificação. Uma função de classificação pode ser implementada como um motor do vetor de suporte, como um classificador de regulação adaptativa, como um classificador de rede neural, e assim por diante. Um motor vetorial de suporte opera encontrando uma hipersuperfície no espaço de possíveis entradas. A hipersuperfície tenta dividir os exemplos positivos dos exemplos negativos pela maximização da distância entre os mais próximos exemplos positivos e negativos em relação à hipersuperfície. Isto permite a correta classificação dos dados que são similares, mas não idênticos, aos dados de treinamento. Várias técnicas podem ser usadas para treinar um motor vetorial de suporte. Uma técnica usa um algoritmo de otimização mínima seqüencial que decompõe o grande problema de programação quadrática em uma série de pequenos problemas de programação quadrática que podem ser analiticamente resolvidos (Veja Seqüencial Minimal Optimization, em http://research.microsoft.com/~jplatt/smo.html).

Regulação adaptativa é um processo iterativo que executa múltiplos testes em uma coleção de dados de treinamento. A regulação adaptativa transforma um fraco algoritmo de aprendizado (um algoritmo que desempenha em um nível somente um pouco melhor do que

o acaso) em um forte algoritmo de aprendizado (um algoritmo que exibe uma baixa taxa de erro). O fraco algoritmo de aprendizado é executado em diferentes subconjuntos de dados de treinamento. O algoritmo concentra cada vez mais nestes exemplos nos quais seus predecessores tendiam a mostrar erros. O algoritmo corrige os erros feitos pelos fracos aprendedores anteriores. O algoritmo é adaptativo em virtude de ele ajustar as taxas de erro dos seus predecessores. A regulação adaptativa combina regras grosseiras e moderadamente imprecisas de manuseio para criar um algoritmo de alto desempenho. A regulação adaptativa combina os resultados de cada teste separadamente executado em 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 da unidade e funções de ativação). A busca em espaço de ponderação para um conjunto de ponderações que minimiza a função objetiva é um processo de treinamento. Um modelo de rede neural pode usar uma rede de função de base radial ("RBF") e uma descida de gradiente padrão como sua técnica de busca.

A figura 1 é um diagrama que ilustra um gráfico dos documentos retornados como o resultado da busca de uma consulta. Neste exemplo, o subgráfico 100 representa uma parte dos documentos retornados no resultado da busca. Os nós 101-112 representam 12 documentos do resultado da busca. Os nós 101 e 108 representam documentos rotulados. O documento representado pelo nó 101 foi rotulado com a contagem de relevância de 0,75, e o documento representado pelo nó 106 foi rotulado com a contagem de relevância de 0,6. O sistema de propagação gerou as setas entre os nós usando um algoritmo de vizinho mais próximo. Neste exemplo, os nós 102, 103 e 104 são, cada qual, um dos vizinhos k mais próximos em relação ao nó 101, mas os nós 105-112 não são um dos vizinhos k mais próximos. Então, o sistema de propagação calculou a similaridade entre os nós conectados usando um algoritmo de classificação de similaridade. Por exemplo, o nó 101 está conectado no nó 102 com uma seta com a ponderação de 0,8, que indica a similaridade entre os nós conectados.

A figura 2 é um diagrama de blocos que ilustra componentes do sistema de propagação em uma modalidade. O sistema de propagação 230 é conectado em armazenamentos de documento 210 (por exemplo, locais da Internet) por meio da ligação de comunicações 220 (por exemplo, Internet). O sistema de propagação inclui um componente de coleta de dados de treinamento 231, um armazenamento de dados de treinamento 232 e um índice de documento 233. O índice de documento contém um índice dos documentos (por exemplo, páginas da Internet) nos armazenamentos de documento. O índice de documento pode ser gerado por um esquadrinhador da Internet. O índice de documento pode incluir um vetor

de recurso para cada documento que for usado para treinar uma função de classificação. Os vetores de recurso podem representar muitos diferentes tipos de recursos dos documentos, tais como freqüência de documento invertida, palavras-chaves, tamanho da fonte, e assim por diante. O componente de coleta de dados de treinamento submete consultas a um motor de busca (não mostrado) e recebe documentos que casam com as consultas. O motor de busca pode ser independente do sistema de propagação. Em um caso como este, o sistema de propagação pode gerar vetores de recurso dinamicamente a partir dos resultados da busca. O componente de coleta de dados de treinamento pode solicitar que um usuário rotule a relevância de alguns dos documentos que casam com as consultas. O componente de coleta de dados de treinamento armazena as consultas, os resultados da busca (por exemplo, vetores de recurso) e rótulos no armazenamento de dados de treinamento. O sistema de propagação também inclui um componente de propagação de relevância 235, um componente de construção de gráfico 236, um componente de geração de ponderações para gráfico 237, um componente de normalização de ponderações de gráfico 238 e um componente de propagação com base no gráfico 239. O componente de propagação de relevância propaga a relevância dos documentos rotulados até os documentos não rotulados que estão armazenados no armazenamento dos dados de treinamento. O componente de propagação de relevância invoca o componente de construção de gráfico para construir um gráfico que inclui setas que representam os documentos de um resultado de busca. Então, o componente de propagação de relevância invoca o componente de geração de ponderações para gráfico para gerar as ponderações iniciais para as setas do gráfico. O componente de propagação de relevância invoca o componente de normalização de ponderações do gráfico para normalizar as ponderações geradas. Então, o componente de propagação de relevância invoca o componente de propagação de relevância com base em gráfico para realizar a propagação de relevância real dos documentos rotulados até os documentos não rotulados. O sistema de propagação também inclui um componente de criação de função de classificação 241 e uma função de classificação 242. A criação da função de classificação usa os dados de treinamento com a relevância propagada para criar uma função de classificação.

O dispositivo de computação no qual o sistema de propagação pode ser implementado pode incluir uma unidade central de processamento, memória, dispositivos de entrada (por exemplo, teclado e dispositivo de apontamento), dispositivos de saída, (por exemplo, dispositivo de exibição) e dispositivo de armazenamento (por exemplo, unidades de disco). A memória e o dispositivo de armazenamento são mídias legíveis por computador que podem conter instruções que implementam o sistema de propagação. Além do mais, as estruturas de dados e estruturas de mensagem podem ser armazenadas ou transmitidas por meio de uma mídia de transmissão de dados, tais como um sinal em uma ligação de comunicações. Várias ligações de comunicações podem ser usadas, tais como a Internet, uma

rede de área local, uma rede de área ampla e uma conexão discada ponto a ponto.

O sistema de propagação pode fornecer serviços a vários sistemas ou dispositivos computacionais, incluindo computadores pessoais, computadores servidores, dispositivos de mão ou portáteis, sistemas multiprocessadores, sistemas com base em microprocessador, dispositivos eletrônicos programáveis pelo cliente, PCs em rede, minicomputadores, computadores de grande porte, ambientes de computação distribuída que incluem qualquer um dos sistemas ou dispositivos expostos, e congêneres.

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

A figura 3 é um fluxograma que ilustra o processamento do componente de criação de função de classificação do sistema de propagação em uma modalidade. O componente de criação de função de classificação coleta dados de treinamento, propaga a relevância dos documentos rotulados até os 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 insere rótulos para um subconjunto de dados de treinamento. No bloco 303, o componente invoca rótulos para um subconjunto dos dados de treinamento. No bloco 303, o componente invoca o componente de propagação de relevância para propagar a relevância dos documentos rotulados até os documentos não rotulados. No bloco 304, o componente treina a função de classificação usando as relevâncias propagadas.

A figura 4 é um fluxograma que ilustra o processamento do componente de propagação de relevância do sistema de propagação em uma modalidade. Ao componente é fornecido dados de treinamento e ele propaga a relevância dos documentos rotulados até os documentos não rotulados. No bloco 401, o componente invoca o componente de construção de gráfico para construir o gráfico inicial que inclui setas. No bloco 402, o componente invoca o componente de geração de ponderações para gráfico para gerar ponderações que indicam a similaridade entre documentos representada pelos nós conectados. No bloco 403, o componente invoca o componente de normalização de ponderações do gráfico para normalizar as ponderações do gráfico. No bloco 404, o componente invoca o componente de propagação de relevância com base em gráfico para realizar a propagação de relevância. Então, o componente retorna.

A figura 8 é um fluxograma que ilustra o processamento do componente de construção de gráfico do sistema de propagação em uma modalidade. O componente cria uma

matriz quadrática com cada linha e coluna representando um documento. Então, o componente identifica e adiciona uma conexão entre cada nó e seus vizinhos k mais próximos (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 foram 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 foram 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, retorna ao bloco 503 para selecionar o próximo documento *j*. No bloco 506, o componente seleciona os 10 documentos *j* com a menor distância para um documento *i* (isto é, os vizinhos mais próximos) e, então, retorna ao bloco 501 para selecionar o próximo documento *i*.

A figura 6 é um fluxograma que ilustra o processamento do componente de geração de ponderações para gráfico do sistema de propagação em uma modalidade. O componente calcula a similaridade entre documentos conectados com base em uma métrica Manhattan. No bloco 601, o componente seleciona o próximo documento *i*. No bloco de decisão 602, se todos os documentos *i* já foram 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 próprio em 0. No bloco 604, o componente seleciona o próximo documento mais próximo *j* (isto é, o documento conectado) em relação o documento selecionado *i*. No bloco de decisão 605, se todos os documentos mais próximos *j* em relação ao documento selecionado *i* já foram selecionados, então, o componente retorna ao 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* em 1. Nos blocos 607-609, o componente retorna calculando a métrica da distância. No bloco 607, o componente seleciona a próxima dimensão *l* do vetor de recurso. No bloco de decisão 608, se todas as dimensões já foram selecionadas, então, o componente retorna ao bloco 604 para selecionar o próximo documento mais próximo *j*, caso contrário, o componente continua no bloco 609. No bloco 609, o componente ajusta a similaridade entre o documento selecionado *i* e o documento selecionado *j* em suas similaridades atuais multiplicado por uma função da distância entre os recursos selecionados *l* do documento selecionado *i* e do documento selecionado *j*, de acordo com a Equação 2. Então, o componente retorna ao bloco 607 para selecionar a próxima dimensão.

A figura 7 é um fluxograma que ilustra o processamento do componente de normalização de ponderações do gráfico do sistema de propagação em uma modalidade. O componente normaliza as ponderações 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 foram 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 da 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 foram selecionadas, então, o componente retorna ao bloco 701 para selecionar a nova linha, caso contrário, o componente continua no bloco 705. No bloco 705, o componente adiciona as ponderações da linha *i* selecionada e da coluna *j* selecionada no elemento diagonal para a linha *i* selecionada. Então, o componente retorna ao bloco 703 para selecionar a próxima coluna *j* para a linha *i* selecionada. No bloco 706, o componente normaliza a matriz de similaridade de acordo com a Equação 3.

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

Embora o assunto em questão tenha sido descrito em linguagem específica para recursos estruturais e/ou atos metodológicos, entende-se que o assunto em questão definido nas reivindicações anexas não é necessariamente limitado em relação aos recursos ou atos específicos supradescritos. Em vez disto, os recursos e atos específicos supradescritos são divulgados como formas de exemplo da implementação das reivindicações. O sistema de propagação pode ser usado para aumentar os resultados da busca. Por exemplo, um motor de busca pode gerar um resultado da busca com base em certos arquivos de documentos. Então, a relevância do documento do resultado da busca pode ser propagada para documentos de um arquivo diferente usando o sistema de propagação. Então, os documentos de diferentes arquivos com a relevância mais alta podem ser adicionados no resultado da busca. O sistema de propagação pode ser usado para propagar a relevância de documentos rotulados com suas relevâncias em relação a uma única consulta até documentos não rotulados (propagação intraconsulta) ou de documentos rotulados com suas relevâncias em relação a múltiplas consultas até documentos não rotulados (propagação interconsulta). O componente de propagação treina o componente de treinamento separadamente para cada consulta com propagação intra-consulta e, simultaneamente, para múltiplas consultas

com propagação interconsulta. Dessa maneira, a invenção não é limitada, exceto como pelas reivindicações anexas.

REIVINDICAÇÕES

1. Sistema para propagar relevância de documentos rotulados até documentos não rotulados, **CARACTERIZADO** pelo fato de que compreende:

um armazenamento de documento (232) que contém representações de documentos, alguns dos documentos sendo rotulados com relevância em relação a uma consulta e outros dos documentos não sendo rotulados em relação à consulta;

um componente gráfico (236) que cria um gráfico dos documentos com os documentos representados como nós sendo conectados por setas que representam similaridade entre os documentos; e

um componente de propagação de relevância (239) que propaga a relevância dos documentos rotulados até os documentos não rotulados com base na similaridade entre os documentos indicada pela similaridade representada pelas setas no gráfico.

2. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de gráfico inclui:

um componente de construção de gráfico que constrói um gráfico no qual nós que representam documentos similares são conectados por meio de setas;

um componente de geração de ponderações que gera ponderações para as setas com base na similaridade dos documentos representada pelos nós conectados; e

um componente de normalização de ponderação que normaliza as ponderações do gráfico.

3. Sistema, de acordo com a reivindicação 2, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico estabelece setas entre nós usando um algoritmo de vizinho mais próximo.

4. Sistema, de acordo com a reivindicação 3, **CARACTERIZADO** pelo fato de que o algoritmo de vizinho mais próximo usa uma métrica de distância Euclidiana.

5. Sistema, de acordo com a reivindicação 3, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico conecta um nó em seus 10 vizinhos mais próximos.

6. Sistema, de acordo com a reivindicação 2, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico estabelece setas entre cada par de nós.

7. Sistema, de acordo com a reivindicação 2, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico estabelece setas entre nós para criar uma árvore de abrangência mínima.

8. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que a relevância dos documentos rotulados é gerada pela busca de documentos relacionados à consulta em um arquivo de documentos, e os documentos não rotulados não são incluídos no arquivo de documentos.

9. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o

componente de propagação de relevância propaga a relevância usando um algoritmo com base em classificação de cópias.

10. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de propagação de relevância propaga relevância de acordo com a seguinte equação:

$$f^\circ = (1-\alpha)(I-\alpha S)^{-1}y$$

em que f° representa um vetor de relevância propagado, S é uma matriz de similaridade, y representa um vetor de relevância inicial e α representa uma taxa de decadência.

11. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o componente de propagação de relevância propaga a relevância de acordo com a seguinte equação:

$$f^\circ = (I + \alpha S + \alpha^2 S^2 + \ldots + \alpha^N S^N)y$$

em que f° representa um vetor de relevância propagada, S é uma matriz de similaridade, y representa um vetor de relevância inicial e α representa uma taxa de decadência, e em que n representa um expoente para qual f° converge em uma solução.

12. Sistema para propagar relevância de páginas rotuladas em relação a uma consulta até páginas não rotuladas em relação à consulta, **CARACTERIZADO** pelo fato de que compreende:

um armazenamento de página (232) que contém representações das páginas, algumas das páginas sendo rotuladas com relevância em relação a uma consulta e outras das páginas não sendo rotuladas com relevância em relação à consulta;

um componente gráfico que cria um gráfico das páginas com as páginas representadas como nós conectados por setas que representam similaridade entre as páginas, incluindo:

um componente de construção de gráfico (236) que constrói um gráfico no qual nós que representam páginas similares são conectados por meio de setas; e

um componente de geração de ponderações (237) que gera ponderações para as setas com base na similaridade das páginas representada pelos nós conectados; e

um componente de propagação de relevância (239) que propaga a relevância das páginas rotuladas até as páginas não rotuladas com base na similaridade entre as páginas indicada pela similaridade representada pelas setas do gráfico e com base em um algoritmo de classificação de cópias.

13. Sistema, de acordo com a reivindicação 12, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico estabelece setas entre nós usando um algoritmo de vizinho mais próximo.

14. Sistema, de acordo com a reivindicação 13, **CARACTERIZADO** pelo fato de que o algoritmo de vizinho mais próximo usa uma métrica de distância Euclidiana.

15. Sistema, de acordo com a reivindicação 13, **CARACTERIZADO** pelo fato de que o componente de construção de gráfico conecta um nó aos seus 10 vizinhos mais próximos.

16. Sistema, de acordo com a reivindicação 12, **CARACTERIZADO** pelo fato de que o componente de geração de ponderações usa uma métrica de distância Manhattan para representar a similaridade entre as páginas.

17. Sistema, de acordo com a reivindicação 12, **CARACTERIZADO** pelo fato de que cada página é representada por um vetor de recurso e a similaridade entre as páginas é representada pela distância no espaço do vetor de recurso.

18. Mídia legível por computador, **CARACTERIZADA** pelo fato de que contém instruções para controlar um sistema de computador para propagar relevância dos documentos em relação a uma consulta até outros documentos por um método que compreende:

criar (236) um gráfico dos documentos representado como nós conectados por setas com ponderações que representam similaridade entre documentos; e

propagar (239) a relevância dos documentos rotulados até os documentos não rotulados com base nas ponderações das setas entre os nós usando um algoritmo com base em classificação de cópias.

19. Mídia legível por computador, de acordo com a reivindicação 18, **CARACTERIZADA** pelo fato de que a propagação de relevância dos documentos rotulados inclui usar uma expansão Taylor para resolver iterativamente a seguinte equação:

$f^{\circ} = (1-\alpha)(I-\alpha S)^{-1}y$

20. Mídia legível por computador, de acordo com a reivindicação 18, **CARACTERIZADA** pelo fato de que a criação do gráfico inclui conectar setas usando um algoritmo de vizinho mais próximo e estabelecer a ponderação de uma seta com base na distância entre os documentos representara pelos nós conectados pela seta.

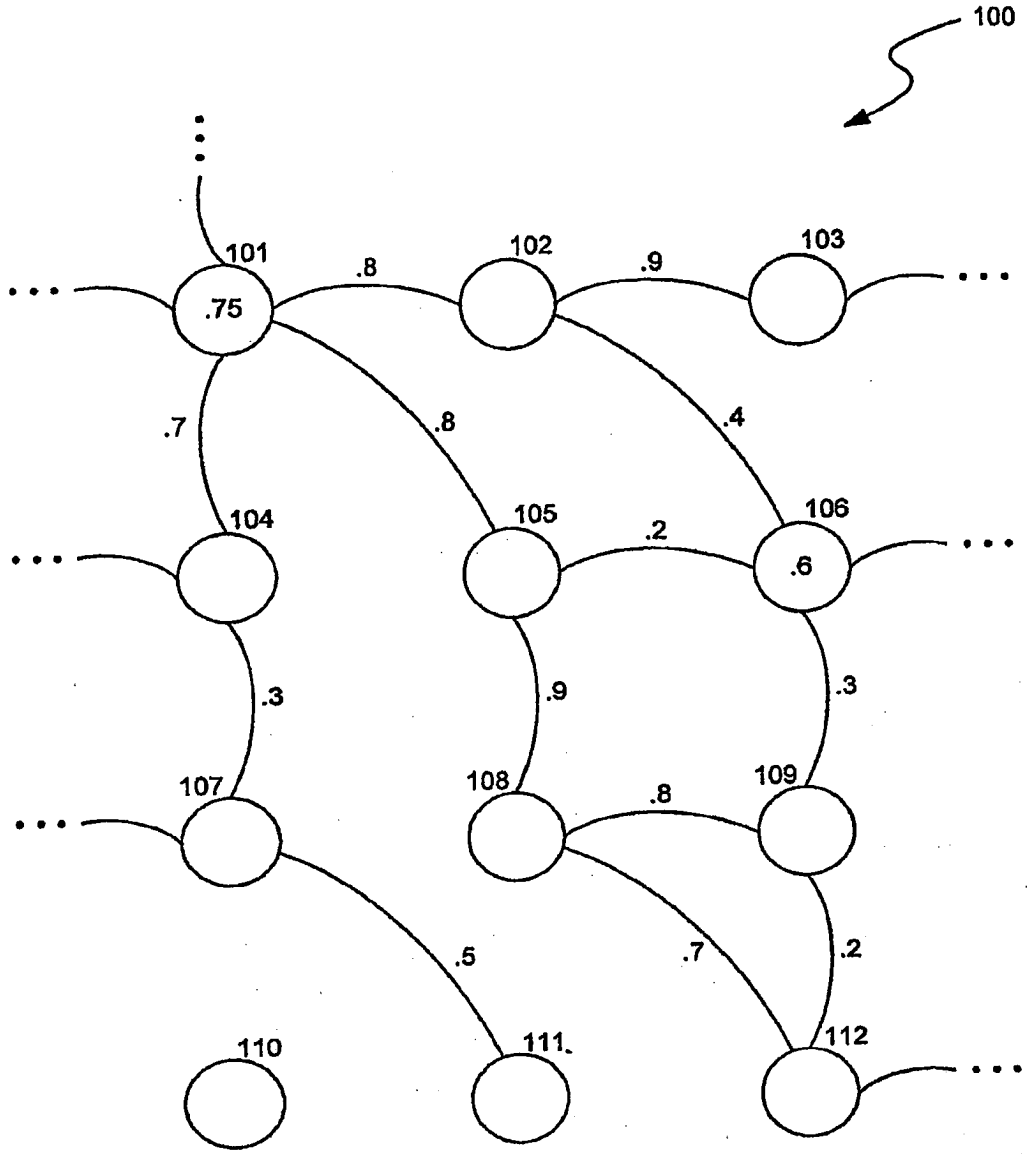

100
101
.75
.8
102
.9
103
.7
.8
.4
104
105
.2
106
.6
.3
.9
.3
107
108
.8
109
.5
.7
.2
110
111
112

FIG. 1

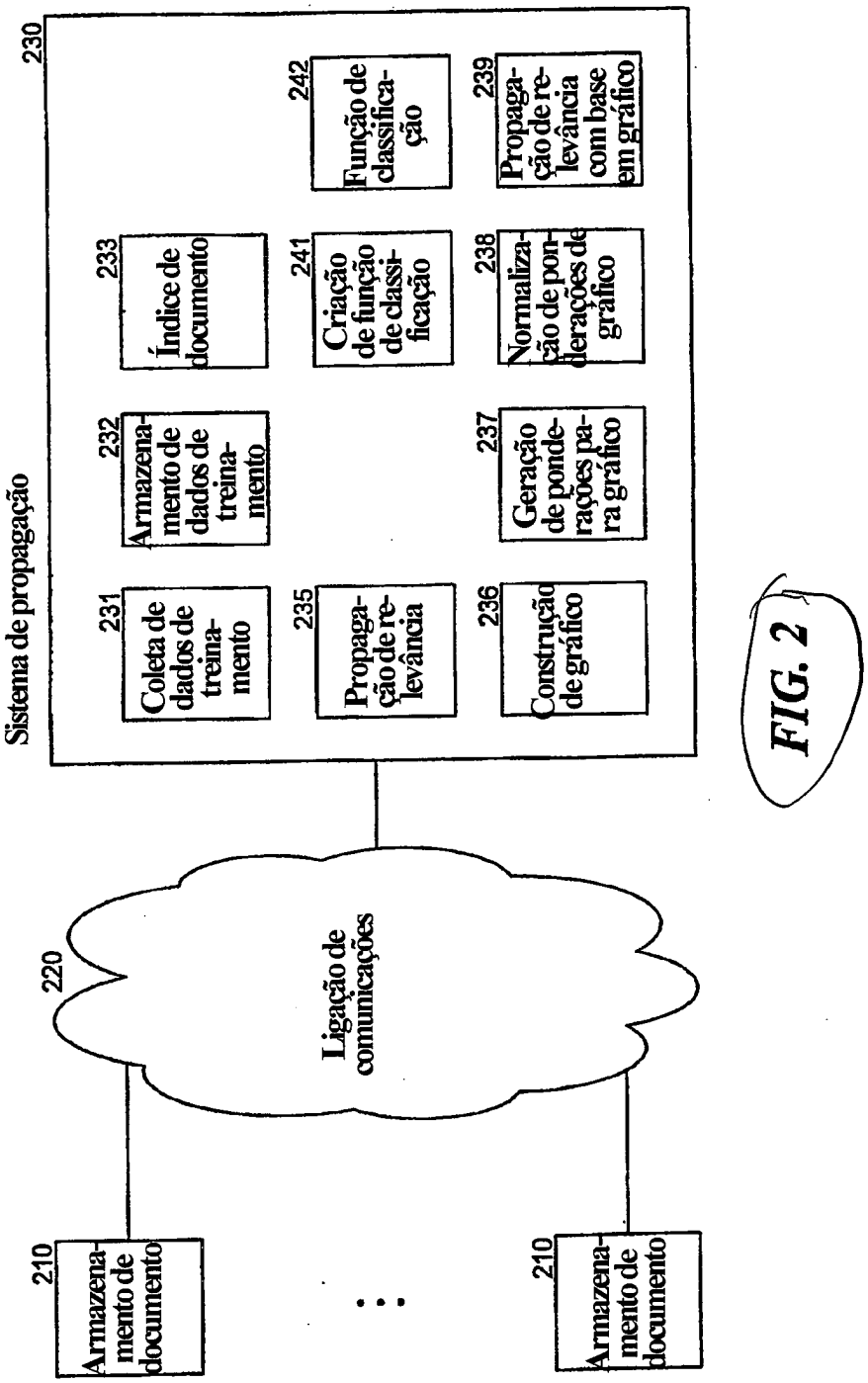

230
Sistema de propagação
231
Coleta de dados de treinamento
232
Armazenamento de dados de treinamento
233
Índice de documento
235
Propagação de relevância
236
Construção de gráfico
237
Geração de ponderações para gráfico
238
Normalização de ponderações de gráfico
239
Propagação de relevância com base em gráfico
241
Criação de função de classificação
242
Função de classificação
220
Ligação de comunicações
210
Armazenamento de documento
210
Armazenamento de documento

FIG. 2

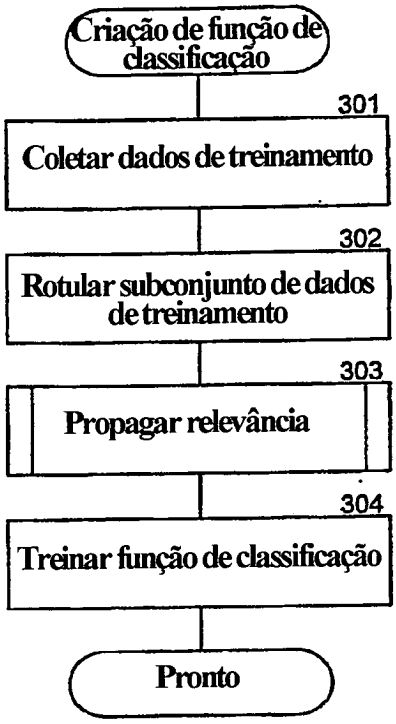

Criação de função de classificação
301
Coletar dados de treinamento
302
Rotular subconjunto de dados de treinamento
303
Propagar relevância
304
Treinar função de classificação
Pronto

FIG. 3

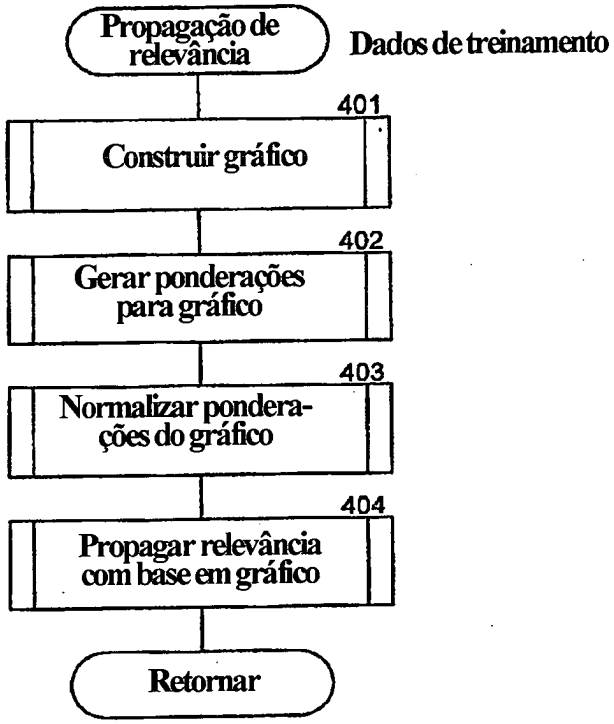

Propagação de relevância
Dados de treinamento
401
Construir gráfico
402
Gerar ponderações para gráfico
403
Normalizar ponderações do gráfico
404
Propagar relevância com base em gráfico
Retornar

FIG. 4

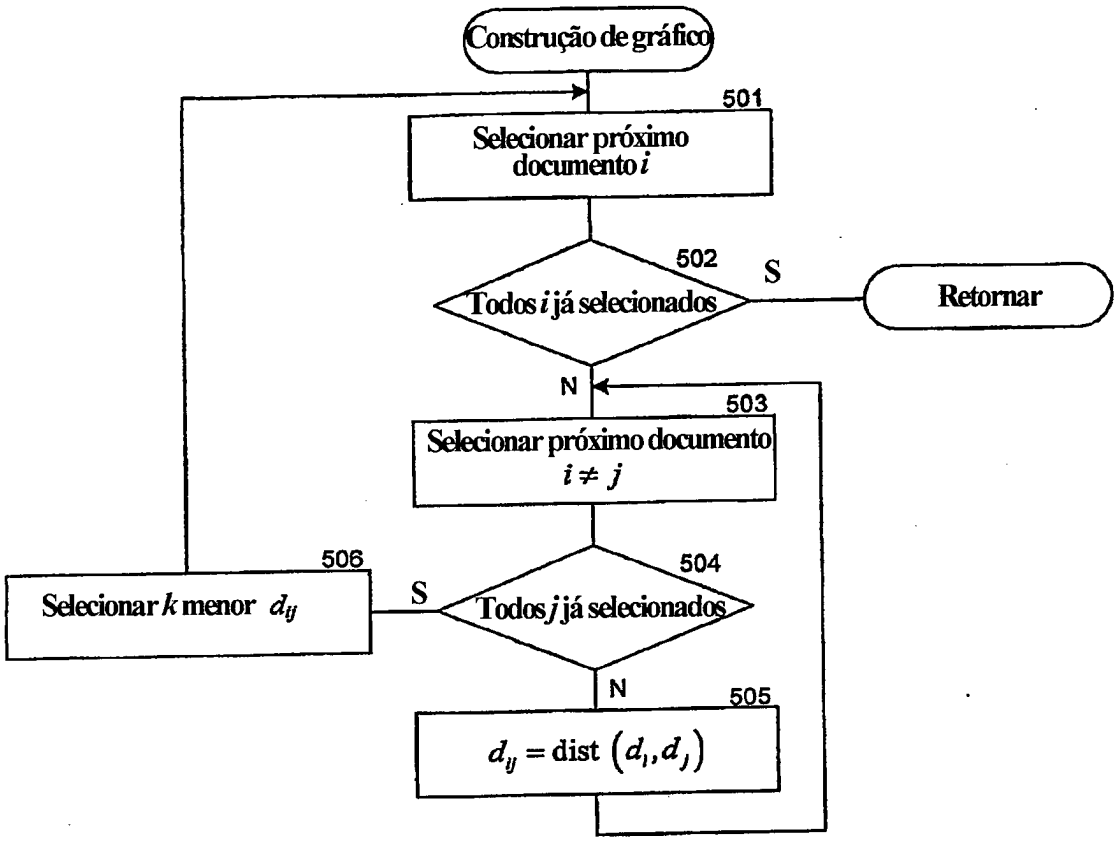

Construção de gráfico
501
Selecionar próximo documento i
502
Todos i já selecionados
S
Retornar
N
503
Selecionar próximo documento i ≠ j
504
Todos j já selecionados
S
506
Selecionar k menor d_{ij}
N
505
$d_{ij} = \text{dist}\left(d_i, d_j\right)$

FIG. 5

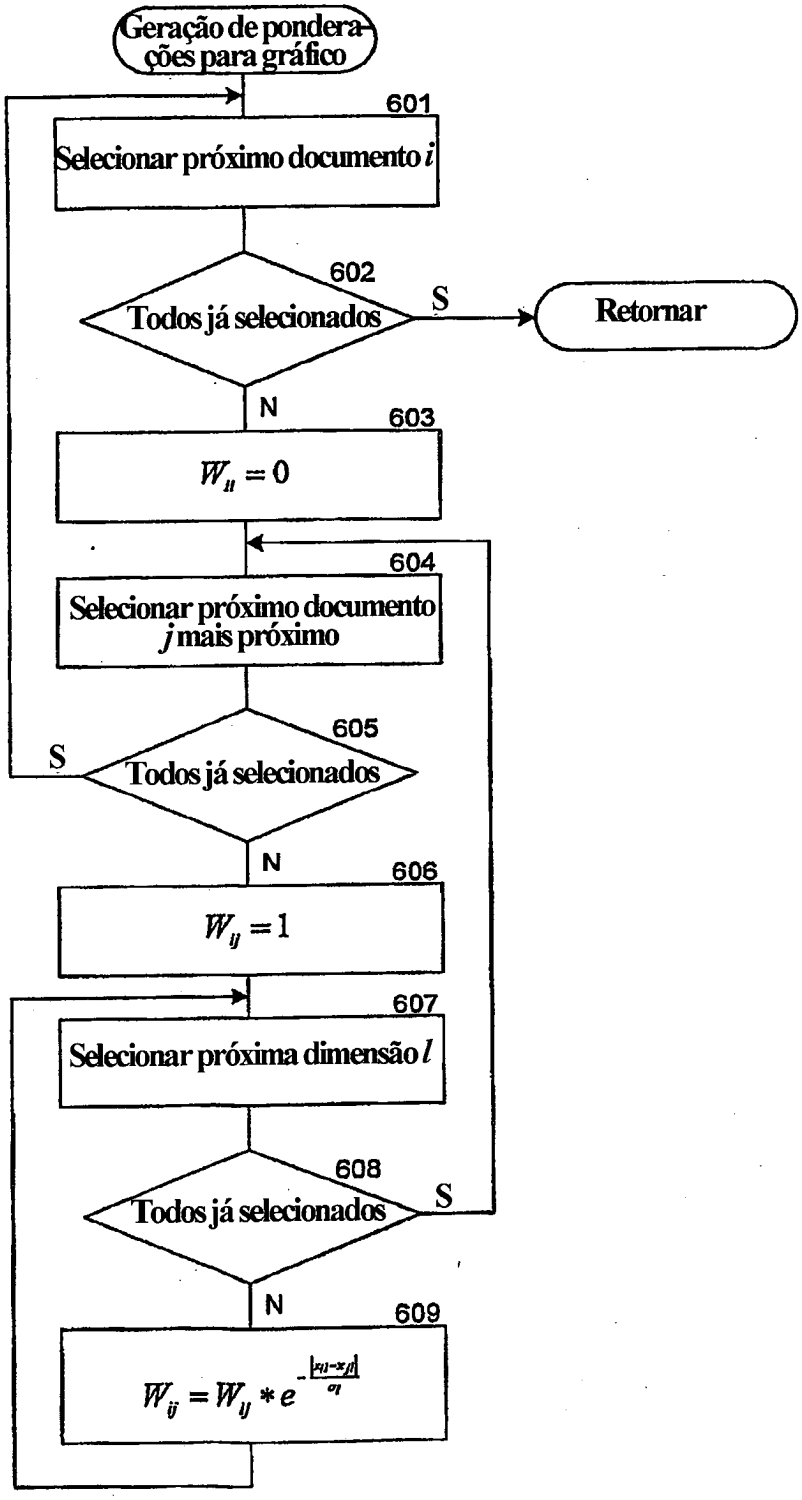

Geração de ponderações para gráfico
601
Selecionar próximo documento i
602
Todos já selecionados
S
Retornar
N
603
$W_{ii} = 0$
604
Selecionar próximo documento j mais próximo
605
S
Todos já selecionados
N
606
$W_{ij} = 1$
607
Selecionar próxima dimensão l
608
Todos já selecionados
S
N
609
$W_{ij} = W_{ij} * e^{-\frac{|x_{il} - x_{jl}|}{\sigma_l}}$

FIG. 6

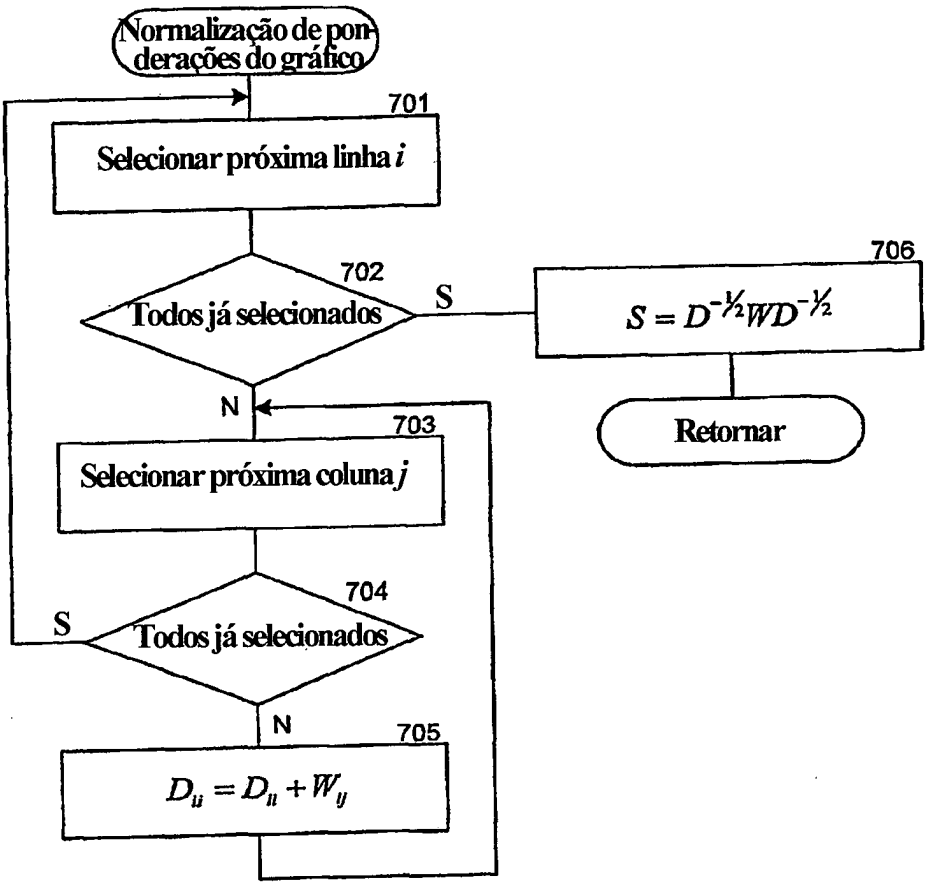

Normalização de ponderações do gráfico
701
Selecionar próxima linha i
702
Todos já selecionados
S
706
$S = D^{-1/2}WD^{-1/2}$
Retornar
N
703
Selecionar próxima coluna j
704
S
Todos já selecionados
N
705
$D_{ii} = D_{ii} + W_{ij}$

FIG. 7

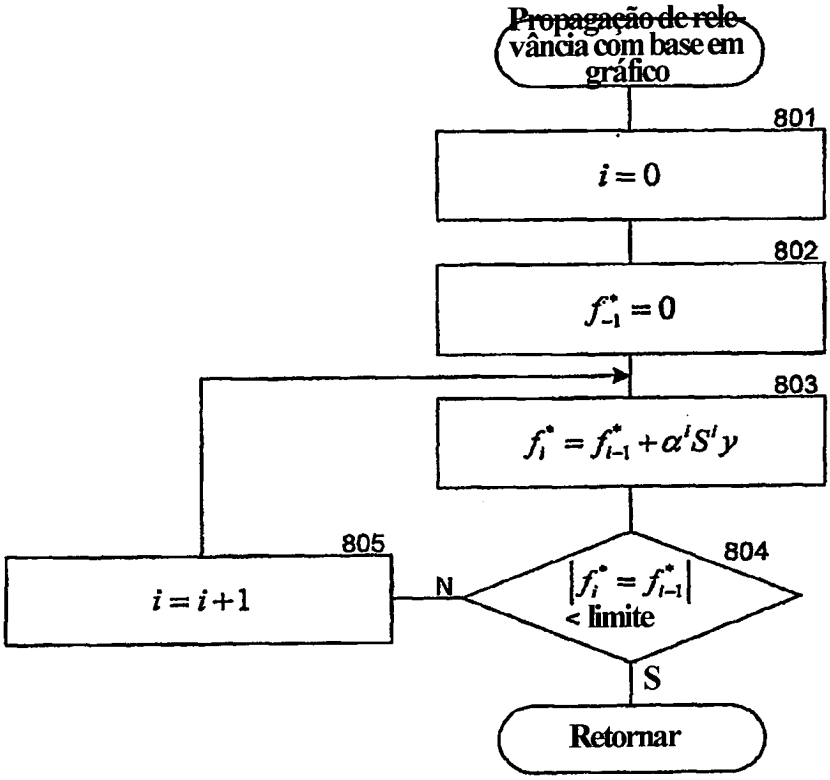

Propagação de relevância com base em gráfico
801
$i = 0$
802
$f_{-1}^* = 0$
803
$f_i^* = f_{i-1}^* + \alpha^i S^i y$
804
$|f_i^* = f_{i-1}^*|$
< limite
N
805
$i = i+1$
S
Retornar

FIG. 8

PI0708074-3

RESUMO

"PROPAGAÇÃO DE RELEVÂNCIA DE DOCUMENTOS ROTULADOS PARA DOCUMENTOS NÃO ROTULADOS"

São fornecidos um método e sistema para propagar a relevância de documentos rotulados em relação a uma consulta até documentos não rotulados. O sistema de propagação fornece dados de treinamento que incluem consultas, documentos rotulados com suas relevâncias em relação às consultas, e documentos não rotulados. Então, o sistema de propagação calcula a similaridade entre pares de documentos nos dados de treinamento. Então, o sistema de propagação propaga a relevância dos documentos rotulados em documentos similares, mas não rotulados. O sistema de propagação pode propagar iterativamente rótulos dos documentos até que os rótulos convirjam em uma solução. Então, os dados de treinamento com as relevâncias propagadas podem ser usados para treinar uma função de classificação.