

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

(22) Data de pedido: 1999.07.08	(73) Titular(es): AMAZON.COM, INC.	
(30) Prioridade(s): 1998.07.15 US 115662	SUITE 1200, 1200 12TH AVENUE SOUTH	
(43) Data de publicação do pedido: 2001.06.06	SEATTLE, WA 98144	US
(45) Data e BPI da concessão: 2012.01.18 063/2012	(72) Inventor(es): RUBEN ERNESTO ORTEGA DWAYNE EDWARD BOWMAN	US US
	(74) Mandatário: ELSA MARIA MARTINS BARREIROS AMARAL CANHÃO RUA DO PATROCÍNIO 94 1399-019 LISBOA	PT

(54) Epígrafe: **SISTEMA E MÉTODO DE CORRECÇÃO DE ERROS DE ORTOGRAFIA EM EXPRESSÕES DE PESQUISA**

(57) Resumo:

UM MOTOR DE PESQUISA UTILIZA CORRELAÇÕES ENTRE TERMOS DE PESQUISA PARA CORRIGIR TERMOS COM ERROS ORTOGRÁFICOS EM EXPRESSÕES DE PESQUISA. AS CORRELAÇÕES SÃO BASEADAS, PELO MENOS EM PARTE, EM SUBMISSÕES DE EXPRESSÕES DE PROCURA HISTÓRICAS AO MOTOR DE PESQUISA. DE UM MODO PREFERIDO, AS CORRELAÇÕES REFLECTEM AS FREQUÊNCIAS COM QUE OS TERMOS DE PESQUISA APARECERAM, HISTORICAMENTE, JUNTOS NA MESMA EXPRESSÃO DE PROCURA E SÃO ARMAZENADOS NUMA TABELA (50) DE CORRELAÇÃO UTILIZANDO LISTAS (62) DE TERMOS RELACIONADOS. NUMA FORMA DE REALIZAÇÃO, UMA TABELA (50) DE CORRELAÇÃO É GERADA PERIODICAMENTE A PARTIR DAS M (E. G., 10) ENTRADAS DIÁRIAS MAIS RECENTES NUM FICHEIRO (36) DE EXPRESSÕES DE PROCURA E REFLECTE, ASSIM, AS PREFERÊNCIAS ACTUAIS DOS UTILIZADORES. EM FUNCIONAMENTO, QUANDO UMA EXPRESSÃO DE PROCURA QUE INCLUI TERMOS DE PESQUISA CORRESPONDENTES E NÃO CORRESPONDENTES É SUBMETIDA AO MOTOR DE PESQUISA, UM PROCESSO (48) DE CORRECÇÃO DE ORTOGRAFIA ACEDE À TABELA (50) DE CORRELAÇÃO PARA GERAR UMA LISTA DE TERMOS QUE SE CONSIDERA ESTAREM RELACIONADOS COM O(S) TERMO(S) CORRESPONDENTE(S). AS GRAFIAS DESSES TERMOS RELACIONADOS SÃO, DEPOIS, COMPARADAS COM A GRAFIA DE CADA TERMO NÃO CORRESPONDENTE UTILIZANDO UMA FUNÇÃO DE COMPARAÇÃO ORTOGRÁFICA QUE COMPARA DUAS CADEIAS DE CARACTERES E GERA UMA PONTUAÇÃO DE SEMELHANÇA. SE SE ENCONTRAR UM TERMO DE SUBSTITUIÇÃO ADEQUADO PARA UM TERMO NÃO CORRESPONDENTE ESPECÍFICO, O TERMO NÃO CORRESPONDENTE É SUBSTITUÍDO PELO TERMO SEMELHANTE RELACIONADO. A EXPRESSÃO DE PROCURA MODIFICADA É, DEPOIS, UTILIZADA PARA REALIZAR A PESQUISA E O UTILIZADOR É NOTIFICADO DA(S) MODIFICAÇÃO(ÕES) FEITAS NA EXPRESSÃO DE PROCURA. NA FORMA DE REALIZAÇÃO DIVULGADA, O MOTOR DE PESQUISA É UTILIZADO NO SÍTIO WEB DE UMA LOJA

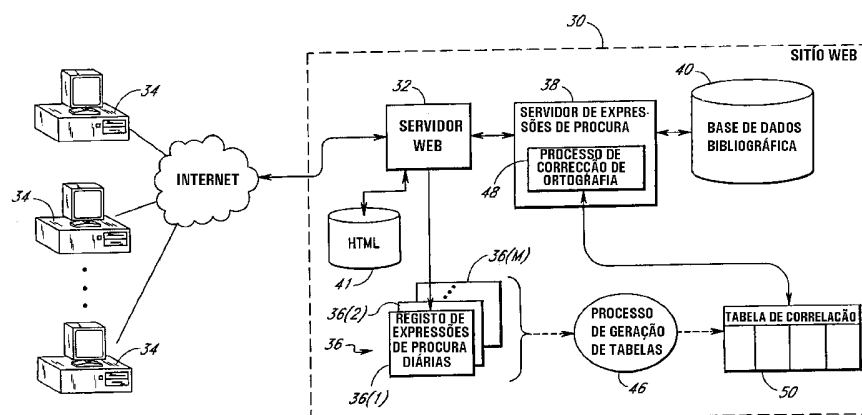
ONLINE PARA AJUDAR OS UTILIZADORES A ENCONTRAR TÍTULOS DE LIVROS, TÍTULOS DE MÚSICAS E OUTROS TIPOS DE PRODUTOS.

RESUMO

"SISTEMA E MÉTODO DE CORRECÇÃO DE ERROS DE ORTOGRAFIA EM EXPRESSÕES DE PESQUISA"

Um motor de pesquisa utiliza correlações entre termos de pesquisa para corrigir termos com erros ortográficos em expressões de pesquisa. As correlações são baseadas, pelo menos em parte, em submissões de expressões de procura históricas ao motor de pesquisa. De um modo preferido, as correlações reflectem as frequências com que os termos de pesquisa apareceram, historicamente, juntos na mesma expressão de procura e são armazenados numa tabela (50) de correlação utilizando listas (62) de termos relacionados. Numa forma de realização, uma tabela (50) de correlação é gerada periodicamente a partir das M (e. g., 10) entradas diárias mais recentes num ficheiro (36) de expressões de procura e reflecte, assim, as preferências actuais dos utilizadores. Em funcionamento, quando uma expressão de procura que inclui termos de pesquisa correspondentes e não correspondentes é submetida ao motor de pesquisa, um processo (48) de correcção de ortografia acede à tabela (50) de correlação para gerar uma lista de termos que se considera estarem relacionados com o(s) termo(s) correspondente(s). As grafias desses termos relacionados são, depois, comparadas com a grafia de cada termo não correspondente utilizando uma função de comparação ortográfica que compara duas cadeias de caracteres e gera uma pontuação de semelhança. Se se encontrar um termo de substituição adequado para um termo não correspondente específico, o termo não correspondente é substituído pelo termo semelhante relacionado. A expressão de procura modificada é, depois, utilizada para realizar a pesquisa e o utilizador é notificado da(s) modificação(ões) feitas na expressão de

procura. Na forma de realização divulgada, o motor de pesquisa é utilizado no sítio Web de uma loja online para ajudar os utilizadores a encontrar títulos de livros, títulos de músicas e outros tipos de produtos.



DESCRIÇÃO

"SISTEMA E MÉTODO DE CORRECÇÃO DE ERROS DE ORTOGRAFIA EM EXPRESSÕES DE PESQUISA"

ANEXO E DIREITOS DE AUTOR

Esta descrição inclui, como anexo, uma listagem C++ de uma função de comparação ortográfica utilizada para comparar duas cadeias de caracteres. O conteúdo do anexo está sujeito a protecção de direitos de autor. O proprietário dos direitos de autor não tem qualquer objecção à reprodução fac-símile do documento de patente ou partes deste, como aparece nos arquivos ou registos do U.S. Patent and Trademark Office ou outro instituto de patentes, mas reserva-se, de outra forma e absolutamente, todos os direitos de qualquer natureza.

CAMPO

A presente invenção refere-se à pesquisa e extracção de informação e, mais especificamente, refere-se a métodos de processamento de expressões de pesquisa.

ANTECEDENTES

Muitos sítios da World Wide Web e serviços online oferecem programas de motores de pesquisa ("motores de pesquisa") para ajudar os utilizadores a encontrarem itens de interesse num

domínio de itens. Por exemplo, sítios Web, tais como AltaVista e Infoseek, proporcionam motores de pesquisa para ajudar os utilizadores a encontrar outros sítios da Web e serviços online, tais como Lexis e Westlaw, implementam motores de pesquisa para auxiliar os utilizadores a encontrarem artigos e opiniões jurídicas. Além disso, comerciantes online proporcionam, normalmente, motores de pesquisa para auxiliar os clientes a encontrar itens a partir de um catálogo online.

Para realizar uma pesquisa utilizando um motor de pesquisa, um utilizador submete uma expressão de procura contendo um ou mais termos de pesquisa. A expressão de procura também pode, explícita ou implicitamente, identificar um campo de registo a pesquisar, tal como o título, autor ou classificação por assunto do item. Por exemplo, um utilizador de um sítio de livraria online pode submeter uma expressão de procura contendo termos que o utilizador pensa que aparecem no título de um livro. Um programa de servidor de expressões de procura do motor de pesquisa processa a expressão de procura para identificar todos os itens que correspondem à expressão de procura. O conjunto de itens identificados pelo programa de servidor de expressões de procura é designado como "resultado de expressão de procura" e é, normalmente, apresentado ao utilizador como uma lista dos itens encontrados. No exemplo da livraria, o resultado da expressão de procura seria, tipicamente, o conjunto de títulos de livros que incluísse todos os termos de pesquisa e seria, normalmente, apresentado ao utilizador como uma listagem hipertextual desses itens.

Quando o utilizador de um motor de pesquisa escreve um termo de pesquisa com um erro ortográfico dentro de uma expressão de procura, tal como por erro de digitação ou por esquecimento do

termo, o termo mal grafado não irá corresponder a qualquer dos termos de banco de dados que são normalmente abrangidos pela pesquisa. Neste caso, muitos motores de pesquisa irão, simplesmente, retornar um resultado de pesquisa nulo (vazio). A apresentação de resultados de pesquisa nulos aos utilizadores, no entanto, pode frustrar significativamente os utilizadores. Para reduzir esse problema, alguns motores de pesquisa ignoram, efectivamente, o(s) termo(s) não correspondente(s) durante a pesquisa. Esta estratégia tem a desvantagem de não ter em conta as informações potencialmente importantes especificadas pelo utilizador e tende a produzir resultados de expressão de procura que contêm um número relativamente grande de itens irrelevantes.

O documento US 5694592 divulga um processo para medir o grau de relevância de uma expressão de pesquisa para documentos particulares, tal como expressão de pesquisa de linguagem natural. Este processo implica a utilização de um léxico semântico (essencialmente, um mapeamento de palavras a categorias semânticas ou significados associados) para avaliar o provável significado de palavras de expressão de procura específicas. O processo também avalia a importância de cada palavra de expressão de procura com base no número de documentos em que a palavra de expressão de procura aparece. Se uma palavra de expressão de procura específica não aparece em nenhum dos documentos que estão a ser pesquisados, mas aparece no léxico, o processo pode substituí-la por um termo relacionado seleccionado do léxico. Num exemplo descrito no documento US 5694592, o léxico é utilizado para substituir a palavra "partida" da expressão de procura, que não aparece em nenhum dos documentos, pela palavra "saída".

Um artigo intitulado "Adapting a Full-text information Retrieval System to the Computer Troubleshooting Domain", publicado pela ACM Press, "Online", Julho de 1994, por P.G. Anick, descreve a afinação de um sistema de extracção de informação para o domínio da resolução de problemas informáticos, onde as expressões de procura por parte do utilizador tendem a ser curtas e os termos de linguagem natural estão misturados com terminologia de uma variedade de sub-linguagens técnicas. Anick menciona, resumidamente, que os dados de registo de expressões de procura podem ser utilizados para detectar erros de ortografia, mas não revela como implementar correcção de ortografia.

SUMÁRIO

A invenção é definida nas reivindicações.

A forma de realização da invenção resolve os problemas anteriores ao proporcionar um sistema e método para corrigir termos mal grafados dentro de expressões de pesquisa. A forma de realização pode incluir uma base de dados de dados de correlação que indicam correlações entre termos de pesquisa. Os dados de correlação podem ser baseados nas frequências com que termos de pesquisa específicos têm, historicamente, aparecido em conjunto dentro da mesma expressão de procura e podem, por exemplo, ser gerados a partir de um ficheiro de registo de expressões de procura. No exemplo, cada entrada na base de dados (implementada como uma tabela) compreende uma palavra-chave e uma lista de "termos relacionados", em que a lista de termos relacionados é composta por termos que apareceram em combinação com a palavra-chave com o maior grau de frequência.

O método de correcção de ortografia pode ser invocado quando se submete uma expressão de pesquisa que inclui, pelo menos, um termo correspondente e, pelo menos, um termo não correspondente. Utilizando o banco de dados de correlação, pode gerar-se, inicialmente, uma lista de termos que se considera estar relacionada com o termo ou termos correspondentes. Isto pode ser conseguido, por exemplo, extraíndo a lista de termos relacionados para cada termo correspondente e se a expressão de procura incluir múltiplos termos correspondentes, combinando estas listas numa única lista de termos relacionados.

Os termos relacionados podem ser comparados, em termos de ortografia, com o(s) termo(s) não correspondente(s) para identificar eventuais substituições adequadas. As comparações de grafia podem ser realizadas utilizando uma função de comparação ortográfica tipo anagrama que gera uma pontuação que indica o grau de semelhança entre duas cadeias de caracteres. Num exemplo, se se encontrar um termo relacionado com uma grafia suficientemente semelhante a um termo não correspondente, o termo não correspondente pode ser automaticamente substituído pelo termo relacionado. O utilizador poderá, em alternativa, ser solicitado a seleccionar o(s) termo(s) de substituição a partir de uma lista de termos. Depois de o termo ou termos não correspondentes terem sido substituídos, a expressão de procura modificada pode ser utilizada para realizar a pesquisa. O utilizador também pode ser notificado da(s) modificação(ões) feita(s) na expressão de procura.

Um benefício importante do método de correcção de ortografia acima descrito em relação aos métodos de correcção de ortografia convencionais é que os termos de substituição seleccionados são consideravelmente mais prováveis de serem os termos que foram

introduzidos pelo utilizador. Este benefício pode ser ainda mais melhorado através da utilização de dados de correlação de termos de pesquisa e, em particular, dados de correlação que reflectem submissões de expressões de procura históricas. O método aumenta a probabilidade de os resultados das expressões de procura conterem itens que são relevantes para o utilizador. Outro benefício é que o método é o adequado para corrigir termos que não constam do dicionário, tais como nomes próprios de autores e artistas e termos extravagantes dentro de títulos e nomes de produtos.

Na forma de realização, podem gerar-se dados de correlação de modo a reflectirem, fortemente, submissões de expressões de procura recentes, reflectindo, assim, profundamente, as preferências actuais dos utilizadores. Isto pode ser conseguido, por exemplo, ao gerar, periodicamente, uma tabela de correlação a partir de um número desejado (e. g., 12) dos registos de expressões de procura diárias mais recentes. A utilização de dados de correlação que reflectem, fortemente, submissões de expressões de procura recentes aumenta ainda mais a probabilidade de as substituições feitas pelo processo de correcção de ortografia serem as pretendidas pelos utilizadores.

BREVE DESCRIÇÃO DOS DESENHOS

Estas e outras características da invenção serão, agora, descritas com referência aos desenhos resumidos a seguir. Estes desenhos e a descrição associada são proporcionados para ilustrar a forma de realização da invenção e não limitar o âmbito da invenção.

A Figura 1 ilustra um sítio Web que implementa um motor de pesquisa de acordo com a invenção.

A Figura 2 ilustra uma página de pesquisa de livros do sítio Web.

A Figura 3 ilustra o formato geral da tabela de correlação da Figura 1.

A Figura 4 ilustra os passos realizados pelo servidor de expressões de procura para processar uma submissão de expressões de procura.

A Figura 5 ilustra o formato geral do registo de expressões da Figura 1.

A Figura 6 ilustra um processo utilizado para gerar, periodicamente, a tabela de correlação a partir do registo de expressões.

DESCRIÇÃO PORMENORIZADA DE FORMAS DE REALIZAÇÃO PREFERIDAS

A presente invenção proporciona um método para corrigir erros de ortografia em expressões de procura que são submetidas a motores de pesquisa. Resumidamente, o método implica utilizar dados de correlação de termos de pesquisa para identificar termos de pesquisa que estão relacionados com o(s) termo(s) de pesquisa grafado(s) correctamente da expressão de procura e avaliar se qualquer destes termos relacionados tem uma grafia semelhante ao(s) termo(s) de pesquisa grafado(s) incorrectamente. Os dados de correlação de termos de pesquisa

são, de um modo preferido, baseados em submissões de expressões de procura históricas e, mais especificamente, nas frequências com que os termos de pesquisa já ocorreram anteriormente em conjunto na mesma expressão de procura. O método pode ser implementado em qualquer um de uma variedade de diferentes tipos de motores de pesquisa, incluindo, por exemplo, motores de pesquisa na Internet, motores de pesquisa de assuntos jurídicos e motores de pesquisa fornecidos por comerciantes online.

A título de ilustração, o método é aqui descrito, no contexto de um motor de pesquisa que é utilizado para auxiliar os clientes da Amazon.com Inc. a encontrar itens (livros, CD, etc.) a partir de um catálogo online de produtos. Ao longo da descrição, será feita referência a vários pormenores específicos de implementação da implementação da Amazon.com. Estes pormenores são proporcionados de modo a ilustrar, em profundidade, a forma de realização da invenção e não a limitar o âmbito da invenção. O âmbito da invenção é definido nas reivindicações anexas.

I. Visão Global de Sítio Web e Motor de Pesquisa

A Figura 1 ilustra o sítio 30 Web da Amazon.com, incluindo componentes utilizados para implementar um motor de pesquisa de acordo com a invenção. Como se sabe na técnica do comércio na Internet, o sítio da Web da Amazon.com inclui funcionalidades para permitir aos utilizadores pesquisar, navegar e fazer compras a partir de um catálogo online de títulos de livros, títulos de músicas e outros tipos de itens. Dado que o catálogo contém milhões de itens navegáveis, é importante que o sítio

proporcione um mecanismo eficiente para auxiliar os utilizadores a encontrar itens.

Como ilustrado pela Figura 1, o sítio 30 Web inclui uma aplicação 32 de servidor Web ("servidor Web"), que processa os pedidos recebidos através da Internet provenientes dos computadores 34 de utilizadores. Estes pedidos incluem expressões de pesquisa que são submetidos por utilizadores para pesquisar produtos no catálogo. O servidor 32 Web regista transacções de utilizadores, incluindo submissões de expressões de procura, num registo 36 de expressões de procura. Na forma de realização representada na Figura 1, o registo 36 de expressões de procura consiste numa sequência de ficheiros 36 de registo de expressões de procura diárias, representando cada ficheiro um dia de transacções.

O sítio 30 Web também inclui um servidor 38 de expressões de procura que processa as expressões de pesquisa, efectuando a pesquisa numa base 40 de dados bibliográfica. A base 40 de dados bibliográfica inclui informações sobre os vários itens que estão disponíveis para compra no sítio. Esta informação inclui, por exemplo, os títulos, autores, editores, descrições de temas e ISBN (Número Internacional Normalizado de Livros) de títulos de livros, e os títulos, artistas, companhias discográficas e classificações de música de títulos musicais. As informações para cada item são organizadas dentro de campos (como um campo de "autor" e um campo de "título"), permitindo que a base 40 de dados seja pesquisada de modo restrito por campos. O sítio também inclui uma base 41 de dados de conteúdo HTML (Linguagem de Marcação de Hipertexto) que inclui, entre outras coisas, páginas com informação sobre produtos que mostram e descrevem os vários itens do catálogo.

A Figura 2 ilustra o formato geral de uma página de pesquisa de livros que pode ser utilizada para pesquisar títulos de livros na base 40 de dados bibliográfica. A página inclui campos 42, 43, 44 de título, autor e assunto e controlos associados que permitem que o utilizador inicie pesquisas de títulos de livros restritas a campos. Os utilizadores também podem aceder a uma página de pesquisa de música (não mostrada) para pesquisar títulos de música utilizando os campos de artista, título, e companhia discográfica. Outras áreas do sítio permitem ao utilizador submeter expressões de pesquisa sem limitar os termos de pesquisa aos campos de base de dados específicos.

Quando o utilizador submete uma expressão de pesquisa a partir da página de pesquisa de livros ou outra página do sítio, o servidor 38 de expressões de procura (Figura 1) aplica a expressão de procura à base 40 de dados bibliográfica, tendo em conta as restrições de campo na expressão de procura. Se o resultado da expressão de procura for um único item, a página de informações sobre produtos do item é apresentada ao utilizador. Se o resultado da expressão de procura incluir múltiplos itens, uma lista dos itens é apresentada ao utilizador através de uma ou mais páginas de resultados de pesquisa (não mostradas) que incluem ligações hipertextuais para as páginas de informações de produtos respectivas dos itens.

No caso de expressões de procura com múltiplos termos, o servidor 38 de expressões de procura liga, efectivamente, com E lógicos os termos de pesquisa uns aos outros para realizar a pesquisa. Por exemplo, se o utilizador digitar os termos "Java" e "programação" no campo de título, o servidor 38 de expressões de procura irá pesquisar e retornar uma listagem de todos os itens que têm os dois termos no título. Assim, se qualquer termo

de pesquisa da expressão de procura não produzir uma correspondência (aqui designada como um "termo não correspondente"), a expressão de procura produzirá um resultado de expressões de procura nulo. Neste caso, pode ser apresentado ao utilizador uma lista de itens que se pensa terem "correspondências mais compatíveis".

Embora o motor de pesquisa aqui descrito usa com E lógicos os termos de pesquisa uns aos outros, deve reconhecer-se que a invenção pode ser aplicada a motores de pesquisa que utilizem outros métodos para combinar termos de pesquisa.

II. Visão Global de uma Característica de Correção Ortográfica

De acordo com a invenção, quando um utilizador submete uma expressão de pesquisa de múltiplos termos que inclui termos correspondentes e não correspondentes, um processo 48 de correção ortográfica (Figura 1) do servidor 38 de expressões de procura utiliza o(s) termo(s) correspondente(s), em combinação com dados de correlação de termos de pesquisa, para tentar corrigir a grafia do(s) termo(s) não correspondente(s). Isto é, de um modo preferido, realizado utilizando uma tabela 50 de correlação de termos de pesquisa para identificar termos adicionais que se pensa estarem relacionados com o(s) termo(s) de pesquisa correspondente(s) e, em seguida, comparando as grafias desses termos relacionados com a(s) grafia(s) do(s) termo(s) não correspondente(s). Por exemplo, se um utilizador submeter a expressão de procura

Java APPI,

e "APPI" for um termo não correspondente, o processo 48 de correcção de ortografia utiliza a tabela de correlação para identificar uma lista de termos que se pensa estarem relacionados com "Java". Esta lista poderia ser a seguinte: café, programação, API, gestão, linguagem. Após comparar as grafias desses termos relacionados com a grafia do termo não correspondente, o processo 48 de correcção de ortografia irá identificar "API" como um termo de substituição candidato. As comparações de grafia são realizadas, de um modo preferido, utilizando uma função de comparação ortográfica tipo anagrama que gera uma pontuação que indica o grau de semelhança entre duas cadeias de caracteres.

Os dados contidos na tabela 50 de correlação indicam as relações entre termos de pesquisa e são utilizados para prever, eficazmente, termos de pesquisa que são susceptíveis de aparecer no interior da mesma expressão de procura. A incorporação destas predições no processo de correcção de ortografia aumenta, significativamente, a probabilidade de um termo de substituição específico identificado pelo processo ser o termo que foi pretendido pelo utilizador.

Se o processo acima identificar um substituto adequado para um termo não correspondente específico, o termo não correspondente é automaticamente substituído pelo termo relacionado. Se não se encontrar um substituto adequado para um termo não correspondente específico, o termo não correspondente é, de um modo preferido, excluído da expressão de procura. Depois de todos os termos não correspondentes serem substituídos ou excluídos, a expressão de procura modificada é utilizada para

pesquisar a base 40 de dados bibliográfica e o resultado é apresentado ao utilizador. O utilizador também é notificado da(s) modificação(ões) feita(s) na expressão de pesquisa.

Como alternativa à substituição automática do(s) termo(s) não correspondente(s), o utilizador pode ser solicitado a seleccionar o(s) termo(s) de substituição a partir de uma lista de termos de substituição candidatos. Isto é realizado, de um modo preferido, através da apresentação de cada termo de substituição candidato como uma hiperligação respectiva (não mostrado) que pode ser seleccionada pelo utilizador para iniciar a pesquisa modificada; o utilizador pode, assim, seleccionar uma expressão de procura modificada e iniciar uma nova pesquisa com uma única acção.

De acordo com outro aspecto da invenção, a tabela 50 de correlação contém ou reflecte, de um modo preferido, informação histórica sobre as frequências de ocorrência de termos de pesquisa específicos em conjunto dentro da mesma expressão de pesquisa. Como representado na Figura 1, estes dados são, de um modo preferido, extraídos do registo 36 de expressões de procura utilizando um processo 46 de geração de tabela. A incorporação destas informações históricas no processo de correcção de ortografia aumenta ainda mais a probabilidade de um termo de substituição encontrado pelo processo ser o termo que foi pretendido pelo utilizador.

No exemplo (descrito abaixo), o processo 46 de geração de tabela regenera diariamente a tabela 50 de correlação a partir dos M ficheiros 36(1)-36(M) de registo diário mais recentes, em que M é um número fixo, tal como dez ou vinte. Esta abordagem de "janela deslizante", de um modo vantajoso, produz uma tabela de

correlação que se baseia apenas em submissões de expressões de procura recentes e que, portanto, reflecte as preferências actuais dos utilizadores. Por exemplo, se um número relativamente grande de utilizadores tiver procurado o livro *Into Thin Air* de Jon Krakauer na semana passada, as correlações entre os termos "into", "thin", "air" e "Krakauer" serão, provavelmente, correspondentemente elevadas; estas correlações elevadas irão, por sua vez, aumentar a probabilidade de um utilizador que escreva um termo mal grafado (de uma expressão de procura com múltiplos termos), enquanto pesquisa este livro ser direccionado para o livro. Qualquer um de uma variedade de outros tipos de métodos de introdução de tendências pode ser aplicado ao registo para se obter um resultado semelhante, incluindo métodos que aplicam uma maior ponderação a submissões de expressões de procura recentes do que a submissões de expressões de procura antigas.

A Figura 3 ilustra o formato geral da tabela 50 de correlação. No exemplo representado na Figura 3 e descrito aqui em pormenor, as correlações entre termos de pesquisa são baseadas apenas na frequência de ocorrência no interior da mesma expressão de procura. Como descrito abaixo, outros tipos de correlações de termos de pesquisa podem, além disso ou em alternativa, ser utilizados. Além disso, embora a implementação divulgada utilize uma tabela para armazenar os dados de correlação, podem utilizar-se outros tipos de bases de dados.

Como ilustrado na Figura 3, cada entrada na tabela de correlação (3 entradas mostradas) tem, de um modo preferido, dois componentes principais: (1) uma palavra-chave 60 e (2) uma lista 62 de "termos relacionados" para essa palavra-chave. A lista 62 de termos relacionados é uma lista dos N termos (e. g.,

20) de pesquisa que apareceram na mesma expressão de procura como a palavra-chave com o maior grau de frequência e é ordenada de acordo com a frequência. Por exemplo, a entrada para a palavra-chave COSMOS é:

COSMOS: ASTRONOMY, SAGAN, UNIVERSE, SPACE, CARL ...

indicando que ASTRONOMY apareceu em conjunto com COSMOS com o maior grau de frequência; SAGAN apareceu com COSMOS com o segundo grau mais elevado de frequência e assim por diante. Considera-se que cada termo que aparece na parte 62 da lista está relacionado com a palavra-chave 60 correspondente devido à frequência relativamente elevada de ocorrência dos termos dentro da mesma expressão de procura. Na implementação aqui descrita, as palavras-chave e termos relacionados são armazenados na tabela sem ter em conta se as palavras são em maiúsculas ou minúsculas, embora essa informação possa, em alternativa, ser preservada.

Com referência, ainda, à Figura 3, cada termo relacionado e cada palavra-chave 60 dentro da tabela 50 inclui, de um modo preferido, um prefixo de campo de caracteres individuais (não mostrado), que indica o campo 42, 43, 44 de pesquisa ao qual o termo corresponde (com base nos campos de pesquisa nos quais os termos foram inseridos pelos utilizadores). Esses prefixos podem, por exemplo, ser os seguintes: A - autor, T = título, S = assunto, R = artista, L = companhia discográfica. Assim, por exemplo, se a palavra-chave COSMOS na Figura 3 tiver o prefixo "T" e o termo relacionado SAGAN tiver o prefixo "A", isto indicaria que um número relativamente grande de expressões de procura incluindo COSMOS no campo 43 de título foi submetido juntamente com SAGAN no campo 42 de autor. Como descrito abaixo,

os prefixos dos termos relacionados são utilizados pelo processo 48 de correcção de ortografia para filtrar eficazmente os termos não correspondentes de campo a partir das listas de termos relacionados, de modo a que um termo não correspondente num campo de pesquisa específico só seja comparado com termos relacionados do mesmo campo. Assim, por exemplo, um termo não correspondente no campo 42 de autor será comparado apenas com outros termos que tenham sido, historicamente, introduzidos no campo 42 de autor. O processo 48 de correcção de ortografia utiliza, de modo semelhante, os prefixos das palavras-chave 60 para seleccionar entradas de tabela que correspondam aos respectivos campos de pesquisa dos termos correspondentes. Por exemplo, se se receber uma expressão de procura errada que inclui o termo correspondente MOUNTAIN no campo 43 de título, o processo 48 de correcção de ortografia irá procurar uma entrada de tabela com a palavra-chave TMOUNTAIN.

Como representado, ainda, na Figura 3, a tabela 50 de correlação também inclui, de um modo preferido, pontuações 64 de correlação que indicam o número de vezes que cada termo relacionado apareceu em combinação com a palavra-chave. Por exemplo, o termo PROGRAMMING tem uma pontuação de 320 na entrada para JAVA, indicando que JAVA e PROGRAMMING apareceram na mesma expressão de procura (nos campos indicados por seus respectivos prefixos de campo) trezentos e vinte vezes. Como descrito abaixo, as expressões de procura que produziram um resultado de expressões de procura nulo são ignoradas pelo processo 46 de geração de tabela e, portanto, não são reflectidas nas pontuações 64 de correlação. Noutras implementações, as pontuações 64 de correlação podem incorporar outros tipos de correlações. Como descrito abaixo, as pontuações 64 são, de um modo preferido, utilizadas para combinar listas de termos

relacionados quando uma expressão de procura tem múltiplos termos correspondentes.

Em funcionamento, quando o servidor 38 de expressões de procura determina que uma expressão de procura contém um termo de pesquisa correspondente e um não correspondente, o processo 48 de correcção de ortografia acede, inicialmente, à tabela 50 de correlação para extrair a lista 62 de termos relacionados associada. Se a expressão de procura incluir múltiplos termos de pesquisa correspondentes, o processo 48 obtém as listas 62 de termos relacionados para cada termo correspondente e combina estas listas entre si (como descrito abaixo) para gerar uma lista de termos relacionados composta. Como indicado acima, o processo, em seguida, compara as grafias dos termos relacionados resultantes com a(s) grafia(s) do(s) termo(s) não correspondente(s) e substitui ou exclui cada termo não correspondente da expressão de procura. Durante o processo de comparação de grafias, cada termo não correspondente é, de um modo preferido, comparado apenas com os termos relacionados abrangidos pelo campo de pesquisa do termo não correspondente.

Um benefício importante deste método é que é particularmente bem adequado para a correcção de erros ortográficos de termos que não aparecem no dicionário. A técnica é, assim, particularmente útil para a identificação de itens que tendem a ser caracterizados por termos não existentes em dicionários. Estes itens incluem, por exemplo, produtos vendidos por comerciantes online, opiniões jurídicas (normalmente identificadas pelos nomes das partes) e empresas e os seus sítios Web. No contexto de uma loja de livros/música online, por exemplo, o método é útil para corrigir erros ortográficos de nomes próprios de autores/artistas e termos extravagantes que

aparecem nos títulos. Por exemplo, um utilizador pesquisando um livro de Jon Krakauer pode ser capaz de encontrar o livro escrevendo umas poucas palavras do título, mais uma versão mal grafada de "Krakauer." A probabilidade de a pesquisa identificar o livro desejado será particularmente elevada se um número relativamente grande de utilizadores tiver pesquisado recentemente o mesmo livro.

Embora o método utilize, de um modo preferido, correlações de termos de pesquisa que são baseadas em submissões de expressões de procura anteriores, deve compreender-se que os dados de correlação podem, em alternativa, ser gerados a partir de outras fontes. Por exemplo, os dados de correlação podem ser gerados pelo processamento da informação na base 40 de dados bibliográfica para identificar termos que aparecem juntos no mesmo registo de item, título, descrição de item, resenha de livro ou outro campo de base de dados; na realização deste processo, pode conceder-se um peso de ponderação a cada item que é proporcional, por exemplo, ao número de unidades desse item vendidas durante a última semana.

Além disso, o método pode ser modificado para incorporar outros tipos de correlações, incluindo correlações com base em outros tipos de acções de utilizador históricas. Por exemplo, na extracção de dados de correlação do registo 36 de expressões de procura, pode atribuir-se um maior peso a submissões de expressões de procura que, com base nas acções subsequentes dos utilizadores, se considerou terem produzido um resultado de expressões de procura bem sucedido. Este sucesso pode ser presumido, por exemplo, se o utilizador vir, comprar ou adicionar a um carrinho de compras um item encontrado pela pesquisa.

Com referência, ainda, à Figura 1, o servidor 32 Web, servidor 38 de expressões de procura, processo 46 de geração de tabela e software de base de dados são executados num ou mais servidores baseados em Unix e estações de trabalho (não mostradas) do sítio 30 Web. A tabela 50 de correlação é armazenada na RAM (memória de acesso aleatório) na mesma estação de trabalho utilizada para implementar o servidor 38 de expressões de procura.

III. Método de Processamento de Expressões de Procura

O processo de correcção de ortografia será, agora, descrito em maior pormenor com referência à Figura 4, que é um diagrama de fluxo dos passos executados pelo servidor 38 de expressões de procura (Figura 1) quando um utilizador submete uma expressão de procura. Para ilustrar este processo, assume-se que o utilizador está a pesquisar livros sobre caminhadas pela Appalachian Trail e digitou a seguinte expressão de procura no campo 44 de assunto (Figura 2): "hike Appalatian trail." Também se assume que "Appalatian" é um termo não correspondente (embora "Appalachian" seja correspondente) e que "hike" e "trail" são termos correspondentes tendo as seguintes listas de termos relacionados:

HIKE: CAMPING (235), WALKS (160), TRAIL (150)

TRAIL: BIKE (200), APPALACHIAN (165), WALKS (50)

Também se assume que todas as palavras-chave e termos relacionados acima têm um prefixo de campo de "S" para "assunto".

Como indicado pelo passo 70, o servidor 38 de expressões de procura aplica, inicialmente, a expressão de procura à base 40 de dados bibliográfica. Como descrito nos passos 72 e 74, se um ou mais itens forem encontrados, o servidor de expressões de procura devolve uma lista desses itens ao servidor 32 Web. O servidor 32 Web, por sua vez, incorpora esta lista em uma ou mais páginas de resultados de pesquisa ou, se apenas um item for encontrado, devolve a página de informações sobre o produto para esse item. Além disso, o servidor Web regista o número de itens encontrados no registo 36 de expressões de procura (ver Figura 5). No presente exemplo, não seriam encontrados quaisquer itens uma vez que "Appalatian" não existe como uma palavra de assunto na base 40 de dados bibliográfica.

Se, como no presente exemplo, o número de itens encontrados for zero no passo 72, o servidor 38 de expressões de procura determina se a expressão de procura inclui termos correspondentes e não correspondentes (passo 76). Se isso acontecer, o servidor 38 de expressões de procura invoca o seu processo 48 de correcção de ortografia (passos 80-90 na Figura 4) para tentar corrigir o(s) termo(s) não correspondente(s). Se isso não acontecer, uma mensagem é devolvida ao utilizador (passo 78), indicando que não se encontraram correspondências exactas; neste caso, o servidor 38 de expressões de procura também pode gerar e devolver uma lista de "correspondências compatíveis", que pode incluir itens que contêm apenas um subconjunto dos termos correspondentes. O servidor 38 de expressões de procura poderia, além disso ou em alternativa, ser configurado para invocar um método de correcção de ortografia alternativo (não mostrado) para tentar corrigir qualquer(quaisquer) termo(s) não correspondente(s). No presente exemplo, o processo 48 de correcção de ortografia seria invocado

dado que a expressão de procura inclui termos correspondentes e não correspondentes.

O processo de correcção de ortografia começa, no passo 80, por extrair a lista de termos relacionados para cada termo correspondente da tabela 50 de correlação. Se nenhuma lista de termos relacionados (entradas de tabela) for encontrada durante o passo 80, o processo, de um modo preferido, devolve uma mensagem de resultado de expressões de procura nulo e termina (não mostrado). Em alternativa, o processo poderia utilizar um método de correcção de ortografia alternativo para tentar corrigir o(s) termo(s) não correspondente(s) ou poderia gerar e devolver uma lista de "correspondências compatíveis".

Se, como no presente exemplo, a expressão de procura incluir múltiplos termos correspondentes, as listas de termos relacionados para estes termos são combinadas entre si no passo 80 para formar uma lista de termos relacionados composta. As listas podem ser combinadas, por exemplo, através da combinação das listas enquanto se somam as pontuações de correlação de quaisquer termos de intersecção (termos que aparecem em várias listas) e, em seguida, através da ordenação da lista composta desde a pontuação da frequência mais elevada para a mais baixa. No presente exemplo, este método produziria a seguinte lista de termos relacionados: CAMPING (235), WALKS (210), BIKE (200), APPALACHIAN (165), TRAIL (150). A lista composta ordenada pode ser truncada para reduzir a carga de processamento dos passos subsequentes.

O processo de correcção de ortografia, em seguida, entra num ciclo (passos 84-90), em que as grafias do(s) termo(s) não correspondente(s) e dos termos relacionados são comparadas. Em

cada passagem deste ciclo, o processo compara um termo não correspondente com a lista de termos relacionados, um a um (passo 84), excluindo quaisquer termos relacionados tendo prefixos de campo que não correspondam ao campo de pesquisa do termo não correspondente. As comparações são realizadas utilizando uma função de tipo anagrama que compara duas cadeias de caracteres e devolve uma pontuação de semelhança numérica. A pontuação de semelhança indica o grau de semelhança entre as grafias das duas cadeias e, assim, indica a probabilidade de o termo relacionado específico ser um substituto adequado para o actual termo não correspondente. Uma pontuação de semelhança de zero indica o maior grau de semelhança e pontuações cada vez maiores indicam graus de semelhança progressivamente menores.

Os passos executados pela função de comparação ortográfica para comparar `STRING1` a `STRING2` são apresentados abaixo, em que o `RESULTADO` variável representa a pontuação que é devolvida pela função. Uma listagem de código de uma implementação C++ do método (implementado como duas funções separadas) é apenso como um anexo.

Passo 1: Ordenar `STRING1` e `STRING2` em ordem alfabética.

Passo 2: Inicializar `POINTER1` para primeiro carácter de `STRING1` ordenado; `POINTER2` para primeiro carácter de `STRING2` ordenado; e `RESULTADO` para zero.

Passo 3: Comparar respectivos caracteres apontados por `POINTER1` e `POINTER2` sem ter em conta se são maiúsculas ou minúsculas. Se iguais, avançar ambos os ponteiros para o próximo carácter da cadeia de caracteres; senão, avançar ponteiro que aponta

para carácter inferior em ordem alfabética e incrementar RESULTADO em 1.

Passo 4: Repetir o passo 3 até que um ponteiro seja avançado para além do último carácter da sua respectiva cadeia de caracteres.

Passo 5: Se um dos ponteiros continuar a apontar para um carácter, adicionar a RESULTADO, o número de caracteres (incluindo o carácter apontado) que permanecem nessa cadeia de caracteres.

No presente exemplo, o passo 84 resulta na comparação do termo não correspondente APPALATIAN com os termos, CAMPING, WALKS, BIKE, APPALACHIAN e TRAIL. As pontuações geradas por estas comparações são listadas no quadro abaixo.

TERMO RELACIONADO	CADEIA DE CARACTERES ORDENADA COMPARADA COM "AAAAILNPPT"	PONTUAÇÃO DE SEMELHANÇA
CAMPING	ACGIMNP	9
WALKS	AKLSW	11
BIKE	BEIK	12
APPALACHIAN	AAAACHILNPP	3
TRAIL	AILRT	7

QUADRO

No passo 86, as pontuações são avaliadas para determinar se qualquer um dos termos relacionados tem uma grafia suficientemente semelhante para ser um termo de substituição

candidato. Na forma de realização preferida, um termo relacionado é considerado semelhante (e, assim, um termo de substituição candidato) se a sua pontuação for menor ou igual a metade do comprimento do termo não correspondentes (o "limiar de semelhança"). No presente exemplo, o termo "Appalachian" satisfaz o teste de semelhança (dado 3 5) e os restantes termos relacionados não o fazem.

Para aumentar a eficiência da função de comparação ortográfica, o processamento no passo 3 pode ser interrompido depois de o RESULTADO exceder o limiar de semelhança. A eficiência também pode ser aumentada interrompendo o processamento de termos adicionais depois de se identificar um termo com uma pontuação suficientemente baixa (e. g., 0).

Se, pelo menos, um termo relacionado satisfizer o teste de semelhança no passo 86, o termo não correspondente é substituído por um termo relacionado tendo a menor pontuação (passo 88). Se múltiplos termos relacionados partilharem a pontuação mais baixa no passo 88, o termo que cair, em primeiro lugar, na lista de termos relacionados (e, portanto, tem a pontuação de frequência mais elevada) é utilizado como substituto.

Se nenhum dos termos relacionados satisfizer o teste de semelhança no passo 86, o termo não correspondente é excluído da expressão de procura. Em alternativa, uma mensagem de resultado de expressões de procura nulo pode ser, nesta altura, devolvida ao utilizador. Depois de todos os termos não correspondentes terem sido substituídos ou eliminados a pesquisa é reiniciada utilizando a expressão de procura modificada e o resultado é devolvido ao utilizador (passo 94).

Se a pesquisa reiniciada no passo 94 produzir um ou mais itens, os itens são apresentados em conjunto com uma mensagem indicando a(s) modificação(ões) feitas na expressão de procura. No presente exemplo, essa mensagem pode dizer o seguinte:

Não se encontrou nenhum livro que incluísse a palavra "Appalatian" no campo de assunto. No entanto, a sua pesquisa foi reiniciada utilizando o termo "Appalachian" em vez de "Appalatian" e encontraram-se os seguintes títulos:

A página de resultados de pesquisa também exhibe e permite, de um modo preferido, que o utilizador edite a expressão de procura modificada, de modo a que o utilizador possa, efectivamente, rejeitar a(s) substituição(ões) do(s) termo(s) de pesquisa e/ou de outro modo reveja a expressão de procura.

Se a pesquisa reiniciada no passo 94 não produzir quaisquer correspondências exactas, pode ser apresentada ao utilizador uma lista de correspondências mais compatíveis ou este pode ser simplesmente notificado de que não foram encontrados resultados. Como alternativa, podem tentar-se substituições e pesquisas adicionais.

Como se deve compreender pelo que foi exposto, o processo da Figura 4 pode ser alterado em qualquer de uma variedade de modos para atingir um objectivo desejado. Por exemplo, poder-se-ia gerar e utilizar tabelas de correlação separadas para diferentes tipos de itens (e. g., livros vs. música) e/ou diferentes tipos de campos (e. g., título vs. assunto). Além disso, o processo de correcção de ortografia poderia ser utilizado apenas para corrigir erros ortográficos dentro de um campo específico (tal como o campo 42 de autor) ou poderia ser aplicado apenas a

termos não correspondentes que não aparecessem num dicionário de termos.

IV. Geração de Tabela de Correlação

O processo 46 de geração de tabela (Figura 1) será, agora, descrito com referência às Figuras 5 e 6.

O processo 46 de geração de tabela é implementado como um processo off-line que é executado periodicamente, tal como uma vez por dia, para gerar uma nova tabela 50 de correlação. Como descrito acima, o processo gera a tabela a partir dos M ficheiros 36(1)-36(M) de registo de expressões de procura diários mais recentes. Utilizando um M relativamente pequeno (e. g., 5) tende a produzir dados de correlação que reflectem, fortemente, tendências de compra de curto prazo (e. g., novos lançamentos, best-sellers semanais, etc.), ao passo que utilizando um M maior (e. g., 100) tende a produzir uma base de dados mais abrangente. Pode utilizar-se uma abordagem híbrida, em alternativa, em que a tabela é gerada a partir de um grande número de ficheiros de registo, mas em que se atribui aos ficheiros de registo mais recentes um maior peso. Por exemplo, as expressões de procura submetidas durante a última semana podem ser contadas três vezes ao gerar as pontuações 64 de correlação, enquanto expressões de procura submetidas de uma semana a um mês atrás podem ser contadas apenas uma vez.

A Figura 5 ilustra o formato geral dos ficheiros de registo de expressões de procura. Cada entrada no registo (quatro entradas mostradas) inclui informações sobre uma determinada transacção HTTP (Protocolo de Transferência de Hipertexto). Por

exemplo, a entrada 100 indica que às 2:23 AM do dia 13 de Fevereiro de 1998, o utilizador 29384719287 submeteu a expressão de procura {autor = Seagal, título = Human Dynamics} na página de pesquisa de livros e que foram encontrados dois itens que correspondiam à expressão de procura. Os valores ITEMS_FOUND no registo indicam, de um modo preferido, os itens de números que correspondem exactamente à expressão de procura original e, portanto, não reflectem "correspondências compatíveis" ou correspondências resultantes de correcções ortográficas.

A entrada 102 indica que o mesmo utilizador seleccionou um item tendo um ISBN de 1883823064 cerca de vinte segundos depois e que essa selecção foi feita a partir de uma página de resultados de pesquisa (como é evidente a partir da linha HTTP_REFERER"). Outros tipos de acções de utilizador, tais como um pedido para colocar um item num carrinho de compras ou para comprar um item, são igualmente reflectidos nos ficheiros de registo. Como indicado pelo exemplo acima, o percurso de navegação de um utilizador específico pode ser determinado pela comparação de entradas no registo 36.

A Figura 6 ilustra a sequência de passos realizados pelo processo 46 de geração de tabela. Neste exemplo, assume-se que o processo é executado uma vez por dia à meia-noite, logo após o ficheiro de registo diário mais recente ser fechado. Supõe-se também que os M-1 ficheiros de registo diários mais recentes já foram processados utilizando os passos 110-114 do processo para gerar ficheiros de resultados diários respectivos.

No passo 110, o processo analisa o novo ficheiro de registo diário para extrair todas as submissões de expressões de procura para as quais ITEMS_FOUND > 0. Ao ignorar as submissões de

expressões de procura que produziram um resultado de expressões de procura nulo (ITEMS_FOUND = 0) proporcionam-se os importantes benefícios de (1) evitar a adição de termos não correspondentes à tabela de correlação como palavras-chave ou como termos relacionados e (2) excluir, para efeitos de consideração, correlações potencialmente "fracas" entre termos correspondentes.

No passo 112, as entradas extraídas no passo 110 são processadas para correlacionar termos de pesquisa por frequência de ocorrência na mesma expressão de procura. Este passo implica contar, para cada par de termos de pesquisa que apareceram na mesma expressão de procura, pelo menos uma vez, o número de vezes que os dois termos apareceram juntos ao longo do dia. Durante este processo, termos idênticos que foram submetidos em diferentes campos de pesquisa são tratados como termos diferentes. Por exemplo, o termo TRAIL com um prefixo de campo de "T" seria tratado como diferente do TRAIL com um prefixo de "S."

Como indicado acima, qualquer um de uma variedade de outros factores pode ser tomado em consideração durante o processo do passo 112. Por exemplo, uma submissão de expressão de procura específica pode ser contada duas vezes se o utilizador, posteriormente, seleccionar um item a partir da página de resultados de pesquisa e contada uma terceira vez se o utilizador, depois, comprar o item ou adicionar o item a um carrinho de compras pessoais. Dados de correlação extrínsecos também podem ser incorporados no processo. Os resultados do passo 112, que estão na forma geral da tabela de correlação da Figura 3, são guardados como um ficheiro de resultados diário.

No passo 116, os ficheiros de resultados diários criados no passo 114 são combinados com os últimos M-1 ficheiros de resultados diários para produzir a tabela 50 de correlação. Como parte deste processo, as listas de termos relacionados são truncadas com um comprimento fixo de N e os dados de correlação resultantes são armazenados numa estrutura de dados de árvore-B para uma consulta eficiente. A nova tabela 50 de correlação é, depois, escrita na RAM no lugar da tabela de correlação existente.

Embora esta invenção tenha sido descrita em termos de determinados exemplos preferidos, outros exemplos que são evidentes para os especialistas na técnica também estão abrangidos pelo âmbito da presente invenção. Consequentemente, pretende-se que o âmbito da presente invenção seja apenas definido por referência às reivindicações anexas.

Nas reivindicações que se seguem, os caracteres de referência utilizados para indicar passos de processo são proporcionados apenas por conveniência de descrição e não implicam uma ordem particular para a realização dos passos.

ANEXO

```
char *sort_string(char*string_to_sort) {
    qsort(string_to_sort, strlen(string_to_sort),
        sizeof(char), qsort_char_compare);
    return string_to_sort;
}

int score_sorted_strings(char* string1, char* string2) {
    int result = 0;
    int finished = FALSE;
    int compare = 0;

    unsigned char* str1_ptr = (unsigned char*) string1;
    unsigned char* str2_ptr = (unsigned char*) string2;

    while (!finished) {

        -if (*str1_ptr == 0 || *str2_ptr == 0)
            finished = TRUE;

        /* a função de pontuação compara dois caracteres; se
           forem iguais, os dois ponteiros são incrementados
           para avançarem, caso contrário, só o ponteiro menor é
           incrementado e comparado outra vez */

        if (!finished) {
            compare=qsort_char_compare((void*)str1_ptr,(void*)
            str2_ptr);

            if ( compare < 0) {
                str1_ptr++;
            }
        }
    }
}
```

```

        result+ +;
    } else if (compare > 0) {
        str2_ptr + + ;
        result+ +;
    } else if ( compare == 0) {
        str1_ptr++;
        str2_ptr+ +;
    }
}

/* Adicionar à pontuação quaisquer resultados */
while (*str1_ptr++ !=NULL) {
    result + + ;
}
while (*str2_ptr++ != NULL) {
    result+ +;
}
return result;
}

```

Copyright 1998 Amazon.com

Protegido como trabalho não publicado

Lisboa, 16 de Março de 2012

REIVINDICAÇÕES

1. Método de processamento de uma expressão de pesquisa que inclui, pelo menos, um termo de pesquisa mal grafado num sistema (30) informático que implementa um motor (38) de pesquisa que é acessível a utilizadores através de uma rede de computadores, compreendendo o método os seguintes passos implementados informaticamente:

(a) receber a expressão de pesquisa de um utilizador através da rede de computadores, compreendendo a expressão de pesquisa uma pluralidade de termos de pesquisa e sendo dirigida a uma base (40) de dados de informação a pesquisar;

(b) identificar na expressão de pesquisa um termo de pesquisa não correspondente que não produza uma correspondência na base de dados de informação e, pelo menos, um termo de pesquisa correspondente que produza uma correspondência na base de dados de informação;

(c) utilizar dados de correlação de termos de pesquisa para identificar uma pluralidade de termos adicionais que se considera estarem relacionados com o, pelo menos um, termo de pesquisa, sendo os dados de correlação de termos de pesquisa baseado, pelo menos, em submissões de expressões de procura históricas; e

(d) comparar os termos adicionais identificados no passo (c) com o termo não correspondente utilizando uma função de comparação ortográfica para identificar um termo adicional que é um termo de substituição candidato correctamente

grafado para o termo não correspondente, estando a função de comparação ortográfica adaptada:

para comparar uma primeira e segunda cadeias de caracteres, ordenando as primeira e segunda cadeias de caracteres e comparando as primeira e segunda cadeias de caracteres ordenadas carácter a carácter; e

para gerar uma pontuação que indica um grau de semelhança entre as primeira e segunda cadeias de caracteres e

sendo a pontuação comparada a um valor limiar para determinar se um termo correspondente adicional é um termo de substituição candidato.

2. Método da Reivindicação 1, em que os dados de correlação de termos de pesquisa se baseiam, pelo menos, em frequências com que termos de pesquisa ocorreram anteriormente na mesma expressão de procura.
3. Método da Reivindicação 1 ou reivindicação 2, compreendendo, ainda, o passo de processamento de submissões de expressões de procura históricas dentro de um registo para gerar os dados de correlação de termos de pesquisa.
4. Método da Reivindicação 3, em que o passo de processamento de submissões de expressões de procura históricas compreende a aplicação uma função de janela ao registo.
5. Método da Reivindicação 3 ou Reivindicação 4, em que o passo de processamento de submissões de expressões de procura

históricas compreende ignorar expressões de procura que produziram resultados de expressões de procura nulos.

6. Método de qualquer Reivindicação anterior, em que o valor de limiar é dependente de um comprimento do termo não correspondente.

7. Método de qualquer Reivindicação anterior, compreendendo, ainda, os seguintes passos implementados informaticamente:

(e) sem a necessidade de intervenção do utilizador, substituir o termo não correspondente pelo termo de substituição candidato na expressão de pesquisa para gerar uma expressão de pesquisa modificada;

(f) aplicar a expressão de pesquisa modificada à base de dados de informação para realizar uma pesquisa; e

(g) avisar o utilizador do resultado da expressão de procura do passo (f) e da substituição feita no passo (e).

8. Método de qualquer das Reivindicações 1-6, em que o passo (d) compreende identificar uma pluralidade de termos de substituição candidatos correctamente grafados para o termo não correspondente e o método compreende, ainda, apresentar ao utilizador uma lista dos termos de substituição candidatos para selecção.

9. Método da Reivindicação 8, em que o passo de apresentar ao utilizador uma lista compreende apresentando cada termo de substituição candidato numa respectiva hiperligação que pode

ser seleccionada pelo utilizador para iniciar uma pesquisa modificada.

10. Método de qualquer Reivindicação anterior, em que o passo (b) compreende identificar uma pluralidade de termos de pesquisa não correspondentes na expressão de pesquisa e o método compreende repetir o passo (e) para cada um dos termos de pesquisa não correspondentes para identificar um termo de substituição candidato para cada termo de pesquisa não correspondente.
11. Método de qualquer Reivindicação anterior, em que o passo (c) compreende:
 - (c1) para cada um de uma pluralidade de termos de pesquisa correspondentes, identificar uma respectiva lista de termos relacionados; e
 - (c2) combinar as listas de termos relacionados identificadas no passo (c1).
12. Método de qualquer Reivindicação anterior, compreendendo, ainda, o processamento de expressões de pesquisa submetidas ao motor de pesquisa por uma pluralidade de utilizadores durante um período de tempo para gerar os dados (50) de correlação, indicando os dados de correlação correlações entre termos de pesquisa com base em, pelo menos, frequências de ocorrências anteriores dos termos de pesquisa na mesma expressão de pesquisa.
13. Método da Reivindicação 12, em que o referido processamento de expressões de pesquisa compreende actualizar os dados de

correlação, substancialmente em tempo real, à medida que as expressões de pesquisa são recebidas dos utilizadores.

14. Método da Reivindicação 12 ou Reivindicação 13, em que o referido processamento de expressões de pesquisa compreende a análise de um registo que inclui expressões de procura submetidas ao motor de pesquisa.
15. Método da Reivindicação 14, em que o referido processamento de expressões de pesquisa compreende, ainda, aplicar uma função de introdução de tendências com base no tempo ao registo para favorecer submissões recentes de expressões de pesquisa em relação a submissões de expressões de pesquisa antigas.
16. Método da Reivindicação 15, em que a referida aplicação de uma função de introdução de tendências com base no tempo compreende a aplicação de uma função de janela ao registo.
17. Método de qualquer das Reivindicações 12 a 16, em que o referido processamento de expressões de pesquisa compreende ignorar expressões de pesquisa que produzem um resultado de expressões de procura nulo.
18. Método de qualquer das Reivindicações 12 a 17, em que o referido processamento de expressões de pesquisa compreende, ainda, a avaliação de acções de utilizadores após submissão de expressões de procura para identificar expressões de pesquisa consideradas como tendo produzido um resultado bem-sucedido e a ponderação mais forte das expressões de pesquisa que produziram um bom resultado na geração dos dados de correlação.

19. Método de qualquer Reivindicação anterior, em que o sistema informático faz parte de um sítio Web e a base de dados de informação inclui informações sobre produtos que estão disponíveis para compra através do sítio Web.
20. Motor de pesquisa para permitir que utilizadores de uma rede de computadores realizem pesquisas de uma base (40) de dados de itens, compreendendo o motor de pesquisa:

um sistema (30) informático que tem dados de correlação de termos de pesquisa armazenados numa memória do mesmo, indicando os dados (50) de correlação de termos de pesquisa correlações entre termos de pesquisa com base em, pelo menos, anteriores submissões de expressões de procura de utilizadores; e

um servidor (38) de expressões de procura que é executado no sistema informático, estando o servidor de expressões de procura adaptado para pesquisar a base de dados de itens utilizando expressões de pesquisa recebidas de utilizadores, estando o servidor de expressões de procura configurado para processar uma expressão de pesquisa com múltiplos termos que inclui termos correspondentes e não correspondentes ao, pelo menos:

aceder aos dados de correlação de termos de pesquisa para identificar uma pluralidade de termos adicionais que se consideram estarem relacionados com o(s) termo(s) correspondentes(s) da expressão de pesquisa;

comparar as grafias dos termos adicionais com a grafia de, pelo menos, um termo não correspondente da

expressão de pesquisa utilizando uma função de comparação ortográfica para determinar se qualquer um dos termos adicionais é um termo de substituição candidato para o termo não correspondente, estando a função de comparação ortográfica adaptada:

para comparar uma primeira e segunda cadeias de caracteres ordenando as, primeira e segunda, cadeias de caracteres carácter a carácter; e

para gerar uma pontuação que indique um grau de semelhança entre as, primeira e segunda, cadeias de caracteres; e

sendo a pontuação comparada a um valor limiar para determinar se um termo adicional correspondente é um termo de substituição candidato.

21. Motor de pesquisa da Reivindicação 20, em que os dados de correlação de termos de pesquisa se baseiam, pelo menos, em frequências com que termos de pesquisa já ocorreram anteriormente na mesma expressão de procura.
22. Motor de pesquisa da Reivindicação 20 ou Reivindicação 21, em que o servidor está configurado para, sem a necessidade de intervenção do utilizador:

substituir um termo não correspondente por um termo adicional para gerar uma expressão de procura modificada; e

pesquisar a base de dados de itens com a expressão de procura modificada.

23. Programa informático compreendendo elementos de programa informático legíveis por computador ou máquina para configurar um sistema informático para implementar o método de qualquer uma das reivindicações 1 a 19.
24. Suporte de programa informático compreendendo um programa informático de acordo com a reivindicação 23.

Lisboa, 16 de Março de 2012

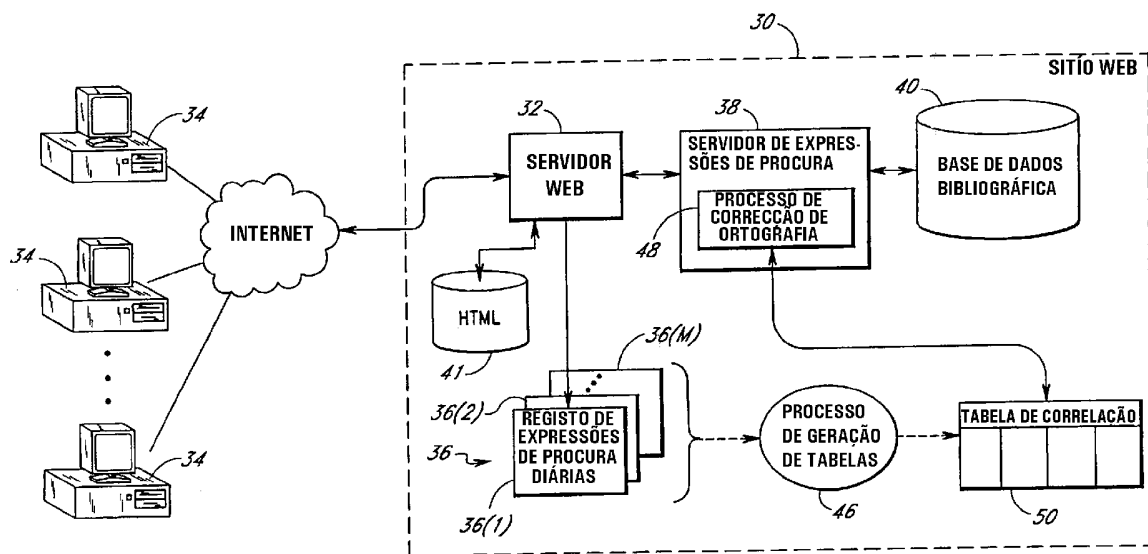


FIG. 1

amazon.com Pesquisa de Livros

Introduzir Autor e/ou Título

Autor: 42

☒ Nome exacto ☐ Último, Primeiro nome ☐ Principio do último nome

Título: 43

☐ Título exacto ☒ Palavra(s) do Título ☐ Principio de Palavra(s) do Título

Dicas para pesquisar por autor / Dicas para pesquisar por título

Pesquisa por Assunto

Assunto: 44

☐ Assunto Exacto ☐ Principio do Assunto ☒ Palavra(s) do Assunto ☐ Principio de Palavra(s) do Assunto

Dicas para Pesquisa por Assunto

Outros Métodos de Pesquisa:

ISBN Editora / Data Pesquisa Rápida

[Pag.inicial Amazon.com](#) | [Procura de Música](#) | [A sua conta](#)

FIG.2

60	COSMOS	JAVA	BIKE	...	N-Terms
	ASTRONOMY (210) SAGAN (180) UNIVERSE (111) SPACE (110) CARL (90) ⋮	PROGRAMMING (320) COFFEE (240) API (120) LANGUAGE (118) MANAGEMENT (60) ⋮	TRAIL (280) REPAIR (190) MOUNTAIN (85) SCHWINN (66) MOAB (19) ⋮	⋮	

FIG.3

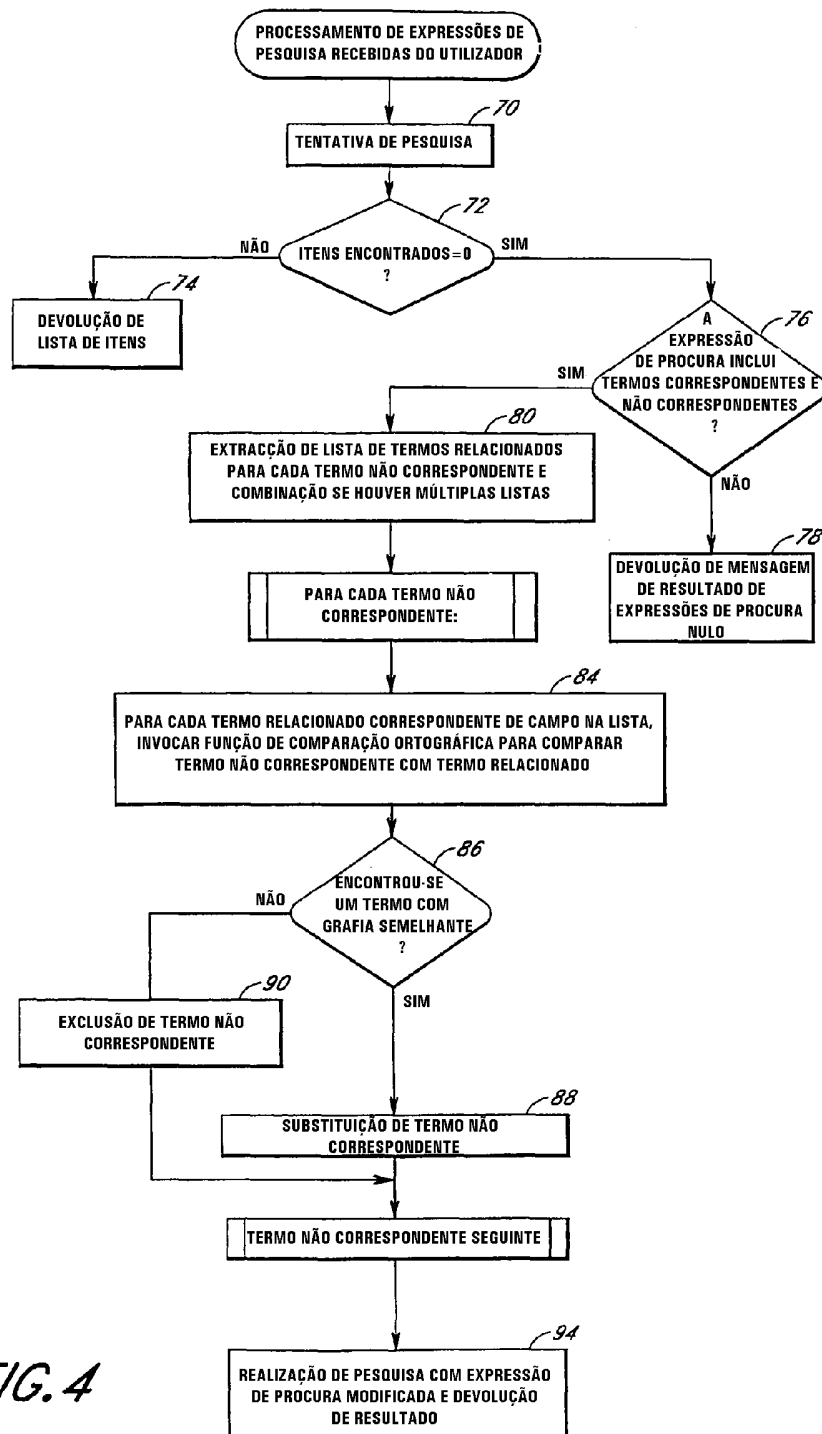
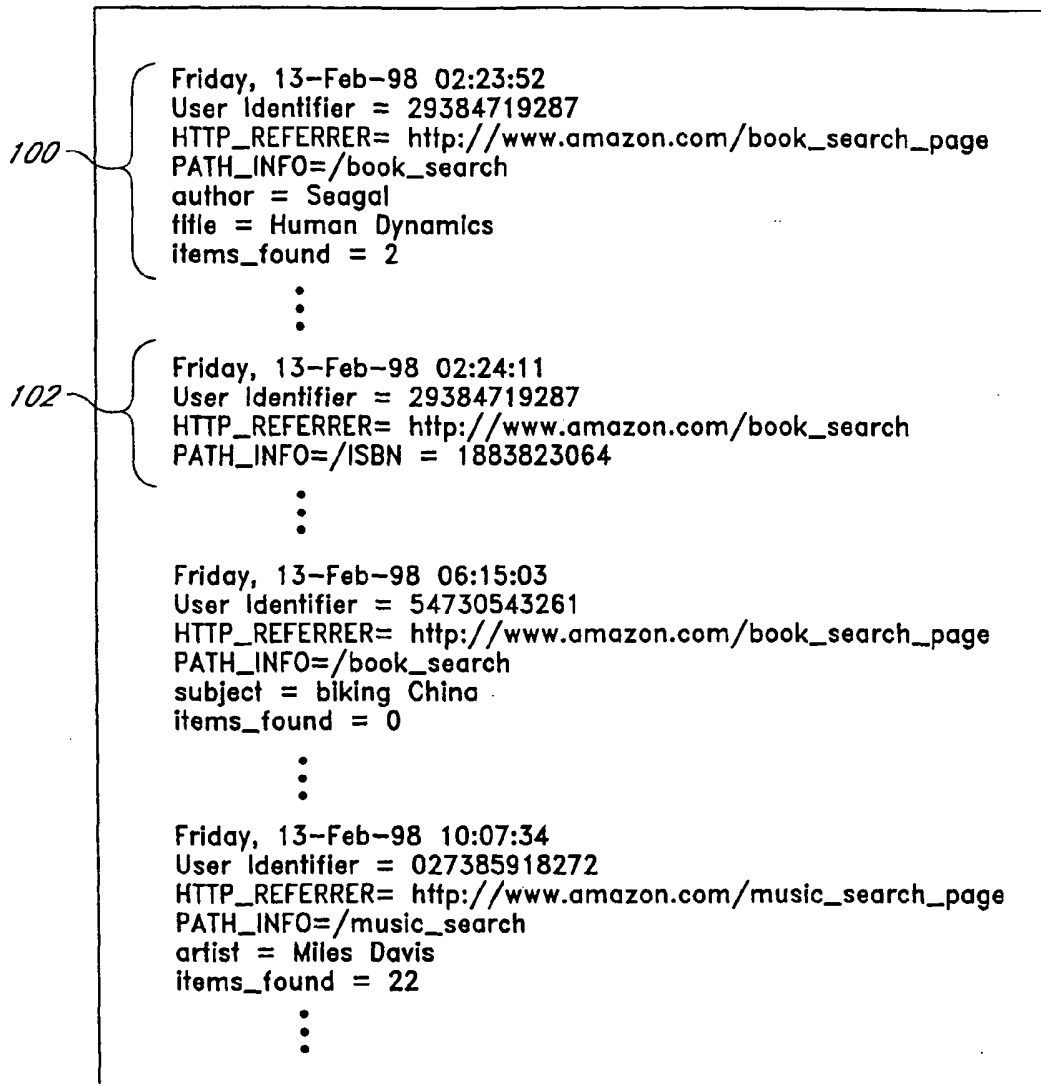
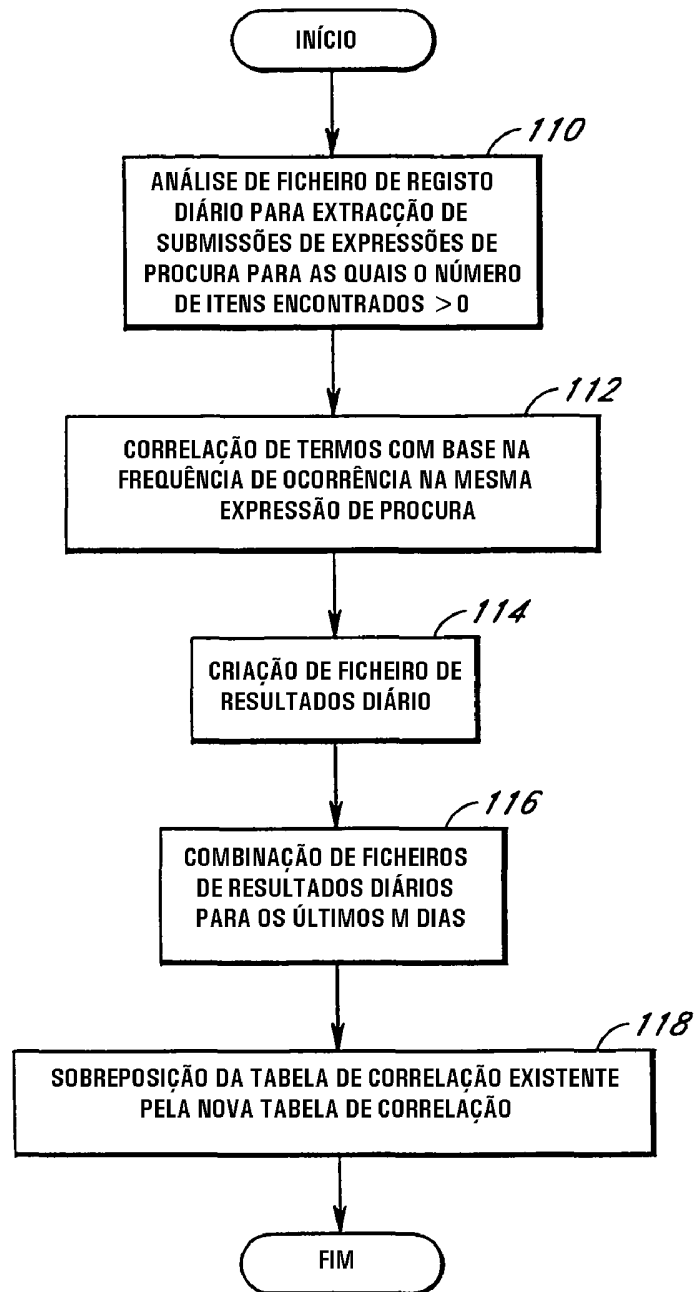


FIG. 4

*FIG. 5*

**FIG. 6**