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

(21) **PI0707800-5 A2**

BRPI0707800A2

(22) Data de Depósito: 15/02/2007
(43) Data da Publicação: 10/05/2011
(RPI 2105)

(51) Int.Cl.:
G06K 9/00
G06K 9/72

(54) Título: **MÉTODO E SISTEMA PARA RESOLVER DADOS DE SAÍDA CONTRADITÓRIOS DE UM SISTEMA DE RECONHECIMENTO ÓPTICO DE CARACTERES**

(30) Prioridade Unionista: 17/02/2006 NO 200607087

(73) Titular(es): Lumex AS

(72) Inventor(es): Hans Christian Meyer, Knut Tharald Fosseide, Mats Stefan Carlin

(74) Procurador(es): Momsen, Leonardos & CIA.

(86) Pedido Internacional: PCT NO2007000053 de 15/02/2007

(87) Publicação Internacional: WO 2007/094684de 23/08/2007

(57) Resumo: METODO E SISTEMA PARA RESOLVER DADOS DE SAÍDA CONTRADITÓRIOS DE UM SISTEMA DE RECONHECIMENTO ÓPTICO DE CARACTERES A presente invenção provê um método e sistema para confirmar palavras duvidosamente reconhecidas como relatadas por um processo de Reconhecimento Óptico de Caracteres usando alternativas ortográficas como argumentos de pesquisa para um utilitário de pesquisa de Internet. O número medido de acertos para cada alternativa ortográfica é usado para prover uma medida de confirmação para a alternativa ortográfica mais provável. Sempre que a medida de confirmação é inconclusa, uma pluralidade de estratégias de pesquisa é usada para alcançar um resultado medido incluindo acertos zero, exceto para uma alternativa ortográfica que é usada como a alternativa correta.

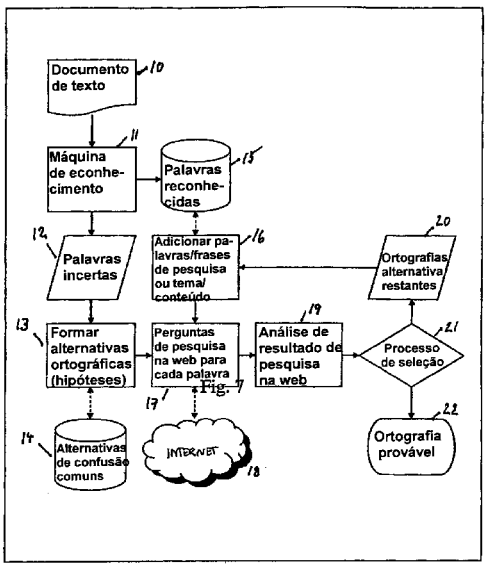


Fig. 7

"MÉTODO E SISTEMA PARA RESOLVER DADOS DE SAÍDA CONTRADITÓRIOS DE UM SISTEMA DE RECONHECIMENTO ÓPTICO DE CARACTERES"

A presente invenção é geralmente relacionada a sistemas de Reconhecimento Óptico de Caracteres (OCR), e especialmente a um método para verificação automática de versão mais provável de palavras duvidosamente reconhecidas como informado pelo processo de reconhecimento.

Existem muitas propostas na técnica anterior para prover reconhecimento óptico de caracteres baseado em imagens de texto. Sistemas de Reconhecimento Óptico de Caracteres (OCR) trabalham bastante bem para documentos de papel varridos de alta qualidade, mas tipicamente falham para varreduras de baixa qualidade ou fontes estranhas. Também há às vezes erros ortográficos nos documentos capturados pelo componente de sistema de OCR. Para ser capaz de re-publicar os documentos, para ser capaz de pesquisar os documentos eletronicamente (registros médicos por exemplo, pesquisa de palavra chave, etc., catálogos eletrônicos, bancos de dados com documentos históricos e informação, etc.), a conversão de imagens de texto para forma executável por computador (converter o texto para texto codificado em ASCII) é um imperativo que proveja um meio para trabalhar com documentos em um modo altamente efetivo em custo, como conhecido a uma pessoa qualificado na técnica. Portanto, há uma necessidade por uma qualidade melhor no resultado de componentes de sistema de OCR para ser capaz de utilizar completamente todas as possibilidades com manipulação de documento eletrônica. A introdução da Internet também foi um fator aumentando demandas para uma qualidade mais alta do processo de OCR como tal. Imagens de texto armazenadas em computadores em formato de PDF por exemplo, são pesquisáveis por navegadores da Internet. Porém, o texto incluído nos arquivos de PDF deve ser convertido para formato digital

legível por computador para ser pesquisável.

Sistemas de software de Reconhecimento Óptico de Caracteres (OCR) podem ser projetados para se adaptar à qualidade de texto e fonte do documento varrido real. OCR adaptável está limitado àqueles caracteres que tem exemplos conhecidos de reconhecimento de caractere robusto, estatísticas conhecidas, e/ou é achado em listas de palavras ou dicionários. Alguns dos caracteres incertos restantes depois do processo de reconhecimento serão caracteres que tanto estão ocorrendo raramente, ou que são facilmente confundidos com outro caractere no processo de reconhecimento provendo um agrupamento de caracteres de interpretações alternativas do caractere. Estes caracteres não podem ser reconhecidos (ou verificados) dentro das estruturas da técnica anterior existentes para OCR. Por exemplo, muitos destes caracteres podem não pertencer a palavras em um dicionário específico de idioma desde que eles podem ser nomes próprios, palavras ou expressões estrangeiras, ou simplesmente sendo de outro idioma. A produção do sistema de OCR geralmente é uma cadeia de caracteres representando o texto como um texto digital. Informação sobre fonte, tamanho e posição também pode ser incluída para ser capaz de recriar o estilo do documento original, por exemplo ao re-publicar o documento. Além disso, a maioria dos sistemas de software de OCR usa uma probabilidade de caractere individual ou valor de contagem para identificar caractere ou palavras duvidosamente reconhecidos, e um verificador ortográfico que provê palavras alternativas para estas palavras duvidosamente reconhecidas.

Na técnica anterior há alguns exemplos de usar a Internet como uma fonte para documentos e informação sobre assuntos, etc., para estabelecer um método para corrigir erros em documentos processados por OCR.

O artigo "Using the Web to Obtain Frequencies for Unseen Bigrams", por Frank Keller e Mirella Lapta, '2003 Association for

Computacional Linguistics', inclui uma investigação e uma abordagem para superar análise de dados para palavras difíceis em um processo de OCR. Uma das questões discutidas neste artigo é se freqüências da Web são adequadas para modelagem probabilística.

O artigo "Text Correction Using Domain Dependent Bigram Models from Web Crawls" por Christoffer Ringsletter et al., AND 2007, descreve como freqüências da web podem ser usadas como um valor de contagem para modificar uma posição existente de candidatos em uma estratégia de correção existente. Nos exemplos descritos no artigo, a Web é usada como um dicionário como conhecido a uma pessoa qualificada na técnica.

O artigo "Precise and Efficient Text Correction using Levenshtein Automata, Dynamic Web Dictionaries and Optimized Correction Models" por Stoyan Mihov et al., Academia Búlgara de Ciências, 2004, descreve um método de construir um dicionário local relacionado ao tema do documento sob processamento de OCR de pesquisas da web. A conclusão é que dicionários locais pequenos provêem o melhor resultado.

Nenhum destes documentos citados da técnica anterior provê um método completo melhorado significativo para corrigir produções de OCR. Portanto, há uma necessidade por uma funcionalidade de OCR avançada que provê confirmação de versão mais provável de palavras duvidosamente reconhecidas em sistemas de OCR.

De acordo com um aspecto da presente invenção, utilitários de pesquisa da Internet podem prover a confirmação só medindo o número de acertos medidos usando uma palavra incerta como um argumento de pesquisa em um utilitário de pesquisa da Internet. De acordo com este aspecto da presente invenção, um argumento de pesquisa provendo zero acertos é considerado como uma certa confirmação que a palavra duvidosamente reconhecida não é esta versão particularmente da palavra sob investigação.

Se o número medido de acertos para uma palavra incerta for muito alto, é certamente possível que esta seja uma versão correta. Porém, de acordo com um aspecto adicional da presente invenção, pesquisas deveriam ser executadas com palavras e/ou combinações alternativas de palavras tal que o número de acertos medidos seja zero para todas as palavras e/ou combinações, exceto para uma palavra e/ou uma combinação. Então a versão mais provável das palavras duvidosamente reconhecidas é esta palavra particular identificada nesta série de medições com uma medição que é não zero.

De acordo com um aspecto da presente invenção, tais etapas de método podem ser implementadas em um programa em um computador interconectado em rede que se comunica com a Internet por uma Interface de Programa Aplicativo (API) se comunicando com sites da Internet. De acordo com este aspecto da presente invenção, o programa implementado recebe entrada sobre palavras duvidosamente reconhecidas de um programa de OCR, executa pesquisas pela API por exemplo, e então mede o número de acertos como informado pelo navegador pelo API. As medições para as alternativas ortográficas diferentes são então usadas para avaliar a palavra mais provável, ou são usadas para iniciar medições adicionais de alternativas ortográficas, usando palavra única, combinação de múltiplas palavras, frases e/ou em combinação com curingas como argumentos de pesquisa adicionais que são medidos.

De acordo com um exemplo de concretização da presente invenção, é possível estabelecer uma medida de confirmação para palavras duvidosamente reconhecidas. Em um exemplo de concretização em que pesquisas da Internet são executadas de acordo com a presente invenção, o número de acertos medidos é todo re-normalizado tal que o número relativo de acertos possa ser comparado. Em concretizações alternativas da presente invenção, medições mais elaboradas e níveis de limiar usados para aceite ou

rejeição de alternativas ortográficas são providas. A medida de confirmação baseada nestes números relativos também pode ser comparada com um limiar de confirmação mais alto e um limiar de confirmação mais baixo. De acordo com este exemplo de concretização, sempre que uma medida de confirmação para uma palavra duvidosamente reconhecida estiver acima do limiar de confirmação mais alto, é considerada como sendo identificada certamente. Se a medida de confirmação estiver abaixo do limiar de confirmação mais baixo, é considerada como sendo não certamente esta versão particular da palavra. Se a medida de confirmação cair entre o limiar de confirmação superior e inferior, investigação adicional da palavra duvidosamente reconhecida é necessária executando pesquisas e medições adicionais.

De acordo com outro aspecto da presente invenção, várias estratégias podem ser usadas para prover alternativas de palavra para a palavra duvidosamente reconhecida, por exemplo, baseado em alternativas para um caractere duvidosamente reconhecido informado por uma função de OCR, estatísticas de letra, etc., e combinando a palavra sob investigação com outras palavras certamente reconhecidas no texto como argumentos de pesquisa. De acordo com um exemplo de concretização da presente invenção, tais palavras e/ou combinações alternativas de palavras são investigadas estabelecendo uma medida de confirmação de acordo com a presente invenção para todos os resultados de pesquisa informados e então usam esta medida como esboçado acima, e pesquisas repetidas com argumentos de pesquisa alternativos até que uma resposta de versão mais provável da palavra sob investigação seja alcançada (tudo zero exceto para uma).

De acordo com outro exemplo de concretização da presente invenção, o limiar de confirmação mais alto e o limiar de confirmação mais baixo podem ser ajustados cooperativamente ou independente um do outro para prover uma afinação dos critérios para categorizar a palavra duvidosamente reconhecida sob investigação.

De acordo com um exemplo de concretização da presente invenção, uma função de OCR relata uma lista de caracteres duvidosamente reconhecidos e as palavras nas quais os caracteres duvidosamente reconhecidos foram encontrados. Além disso, as alternativas que são possível para cada possível versão dos caracteres também são relatadas. Na base destes caracteres alternativos, várias palavras candidatas são criadas como sendo a possível versão correta da palavra, em que cada palavra candidata inclui um dos caracteres alternativos, respectivamente. De acordo com um aspecto da presente invenção, identificar a palavra candidata correta mais provável pode ser alcançado usando cada palavra candidata como um argumento de pesquisa em um utilitário de pesquisa da Internet (usando uma API, por exemplo), e o número medido de acertos de cada palavra forma a base para decidir a versão mais provável da palavra. De acordo com outro exemplo de concretização da presente invenção, a medida de confirmação esboçada acima é usada no processo de decisão.

De acordo com outro exemplo de concretização da presente invenção, sempre que a medição de acertos provê um empate entre candidatas, por exemplo um número igual de acertos entre duas candidatas, as palavras candidatas são combinadas primeiro com a palavra prévia relativa à palavra incerta sob investigação, e então as palavras combinadas são usadas como argumento de pesquisa na Internet, secundariamente a pelo menos uma palavra sucessiva relativa da palavra sob investigação na mesma linha de texto é usada de uma maneira semelhante. Adicionalmente, uma combinação da pelo menos uma palavra prévia, da palavra sob investigação e da pelo menos uma palavra sucessiva também é usada como um argumento de pesquisa. O número de acertos de cada combinação é usado em um processo de confirmação para decidir a versão mais provável das palavras.

De acordo com ainda outro exemplo de concretização da presente invenção, sempre que as combinações de palavras provêem uma

resposta inconclusa, a palavra sob investigação é combinada com uma palavra prévia além da palavra sob investigação. De acordo com o presente exemplo de concretização, a gama de palavras que podem ser selecionadas como uma combinação pode ser limitada a um local a uma distância predefinida, por exemplo tal como 5 palavras da palavra sob investigação. De uma maneira semelhante, as mesmas etapas são executadas com palavras sucessivas, por exemplo, limitado à quinta palavra sucessiva. Porém, qualquer distância da palavra sob investigação pode ser usada, que é uma característica de projeto da presente invenção. De acordo com outra característica de projeto da presente invenção, o local de onde a distância é calculada não precisa ser a própria palavra sob investigação, mas a distância pode ser relacionada a uma área que inclui a palavra sob investigação, por exemplo. Os acertos medidos resultantes destas pesquisas são então usados como uma base para decidir a versão mais provável da palavra.

De acordo com ainda outro exemplo de concretização da presente invenção, as palavras precedentes e as palavras sucessivas que são selecionadas para serem combinadas com a palavra sob investigação não só está baseado em localização relativa para a palavra sob investigação, mas também no número de caracteres que a palavra inclui. De acordo com um aspecto da presente invenção, palavras longas (por exemplo mais de 8 caracteres de comprimento, mas qualquer comprimento pode ser usado e pode ser predefinido ou selecionável por usuário) são preferidas como um qualificador para as palavras sob investigação, como descrito acima.

De acordo com ainda outro exemplo de concretização da presente invenção, a pelo menos uma palavra precedente ou pelo menos uma palavra sucessiva relativa à palavra sob investigação é selecionada na base de freqüência de ocorrência em um idioma específico. Palavras freqüentes são normalmente "palavras pequenas" tais como "e", "o", "em, "de", etc., e podem facilmente ser entendidas como não sendo contribuintes ao processo de

verificação. Portanto, é preferível usar palavras precedentes ou sucessivas com baixa freqüência de ocorrência. Em um exemplo de concretização da presente invenção, o número de ocorrências de uma palavra particular é informado da função de OCR, e um processo de acordo com a presente invenção verifica este número contra um limiar. O número informado de ocorrência e o limiar podem ser re-normalizados como conhecido a uma pessoa qualificada na técnica para prover uma medida relativa de ocorrência.

Porém, palavras com altas freqüências no documento, mas que provêem baixos acertos medidos em pesquisas da Internet, são boas candidatas para uso em pesquisa de combinação com alternativas ortográficas para a palavra sob investigação.

De acordo com ainda outro exemplo de concretização da presente invenção, nomes próprios podem ser reconhecidos como tais em uma base de combinar vários nomes próprios identificados no texto. De acordo com este exemplo de concretização da presente invenção, todas as palavras começando com uma letra maiúscula são tratadas como um nome próprio contanto que o caractere precedente não seja uma marca de pontuação de fim de oração, tal como ".!?:". Combinando pelo menos dois nomes próprios encontrados no texto, o processo de confirmação pode retornar uma resposta correta. De acordo com este exemplo de concretização da presente invenção, a função de OCR relata todos os possíveis candidatos de serem nomes próprios para o processo de confirmação ao executar o processo de reconhecimento.

De acordo com ainda outro aspecto da presente invenção, sistemas de OCR são freqüentemente usados em um contexto específico, por exemplo em um sistema de arquivo em um hospital. Diários de paciente são atualmente freqüentemente registrados e armazenados eletronicamente, mas diários antigos são freqüentemente baseados em papel e precisam portanto ser varridos para serem integrados na versão eletrônica do sistema. De acordo

com um exemplo de concretização da presente invenção, sites da Internet que são usado para a pesquisa no processo de confirmação são selecionáveis. Por exemplo, em um caso com diários de hospital, sites da Internet incluindo informação médica são a melhor escolha para sites a serem pesquisados.

De acordo com outro aspecto da presente invenção, qualquer tipo de conhecimento de contexto relacionado ao documento a ser varrido em um sistema de OCR pode ser usado como qualificadores de palavras. Contexto médico como descrito acima pode ser adicionalmente refinado a especialidades médicas tais como ortopedia, etc. Outros exemplos podem ser história familiar, em que um sobrenome especial é predominante. Outros exemplos podem ser de ciência, agricultura, etc. Comum para todo este "conhecimento" é que é fácil para converter este "conhecimento" em endereços para utilitários de pesquisa incluindo informação pertinente relacionada ao contexto das páginas de documento a serem reconhecidas. Ligações a estas páginas são então usadas ao pesquisar a WEB com as palavras candidatas diferentes de palavras duvidosamente reconhecidas, e os números de acertos para as alternativas diferentes são então usados como uma base para selecionar a palavra mais provável. De acordo com um exemplo de concretização da presente invenção, Agente Professional Copérnico é usado como o utilitário de pesquisa, em que os critério de pesquisa a serem usados são selecionados de acordo com conteúdo das páginas a serem reconhecidas. Neste exemplo de utilitário de pesquisa, é possível selecionar sites de acordo com lei, recursos humanos, governo, ciência, etc.

De acordo com ainda outro aspecto da presente invenção, embora uma palavra seja reconhecida duvidosamente devido a caracteres duvidosamente reconhecidos na palavra, partes de tais palavras ainda podem ser uma palavra reconhecida válida. Por exemplo, "dona de casa" inclui duas palavras "casa" e "dona". Se a parte duvidosamente reconhecida da palavra estiver relacionada com a parte da palavra "dona", pesquisar com

combinações incluindo "casa" simplificaria o processo de confirmação. De acordo com um exemplo de concretização da presente invenção, um dicionário é usado para extrair partes principais identificáveis de palavras duvidosamente reconhecidas. Isto é alcançado tomando a primeira letra da palavra como um argumento para o processo de consulta de dicionário, e então combinando a primeira letra com a próxima letra até que a possível combinação mais longa de letras da palavra que provê um resultado do processo de consulta de dicionário seja identificada. Esta parte da palavra é então usada no processo de pesquisa como um qualificador para o resto da palavra que precisa ser confirmada como a palavra mais provável. Se o resultado do processo de consulta de dicionário estiver inconcluso, o processo continua de acordo com um dos exemplos de concretizações descrito acima.

De acordo com ainda outro aspecto da presente invenção, as mesmas etapas de um método de acordo com a presente invenção podem ser utilizadas em um processo de verificação ortográfica. Algoritmos de verificação de ortográfica na maioria dos casos serão capazes de verificar ortografia dessas palavras que fazem parte do dicionário específico de idioma. Algumas classes de palavras como palavras em idiomas estrangeiros e nomes próprios não podem serem esperadas serem achadas no dicionário específico de idioma como há freqüentemente limitações para o tamanho e consistência do dicionário. Utilizando os aspectos da presente invenção como esboçado acima, um método incluindo as etapas de acordo com a presente invenção pode resolver palavras grafadas incorretas.

De acordo com ainda outro aspecto da presente invenção, palavras duvidosamente reconhecidas são encontradas freqüentemente em sistemas de reconhecimento de fala igualmente. Sempre que um processo de reconhecimento, sendo um processo de reconhecimento óptico ou reconhecimento de fala, etc., relata palavras duvidosamente reconhecidas, possíveis variações da palavra duvidosa é então estabelecida, por exemplo por

sugestões de alternativas de caractere para um caractere duvidosamente reconhecido como proposto pelo próprio processo de reconhecimento, ou identificando palavras reais como parte de uma palavra como descrito acima, pesquisando a WEB pode prover um processo identificando a palavra mais provável como o reconhecimento correto da palavra.

De acordo com ainda outro aspecto da presente invenção, caracteres duvidosamente reconhecidos podem ser combinações de dois ou mais caracteres. Por exemplo, o caractere "m" pode ser uma combinação de "r" e "n" ou outro modo. Quer dizer, um "r" e "n" duvidosamente reconhecido pode ser um "m". Está portanto dentro da extensão da presente invenção prover soluções com número variável de caracteres duvidosamente reconhecidos.

Figura 1 ilustra um exemplo de uma palavra difícil "Helligolav".

Figura 2 ilustra um exemplo de reconhecimento dúbio das letras "N" e "H".

Figura 3 ilustra um retrato de um navio encontrado ao pesquisar a Internet.

Figura 4 ilustra um exemplo de resultado de pesquisa usando as frases de pesquisa "Helligolav" e "Nelligolav".

Figura 5 ilustra outro exemplo de palavra reconhecível difícil.

Figura 6 descreve um fluxograma de um exemplo de método de acordo com a presente invenção.

Figura 7 ilustra um exemplo de saída de um programa de OCR existente.

De acordo com um aspecto da presente invenção, o processo de confirmação é executado em três etapas principais. O processo de reconhecimento, por exemplo um processo de reconhecimento óptico (OCR), primeiro identifica caracteres duvidosamente reconhecidos junto com

alternativas de classificação de caractere para este caractere. Figura 7 ilustra um exemplo de saída de um programa de OCR disponível comercial. Um exemplo do processo de OCR poderia ser que o caractere "i" pode ter as alternativas "l" e "j". Secundariamente, a palavra ou frase que o caractere faz parte é usada como entrada a um utilitário de pesquisa de web formando uma pesquisa para cada combinação de caractere alternativa dessa palavra ou frase particular. Por exemplo, com as alternativas "i","l" e "j", três alternativas são usadas para a palavra sob investigação. Em terceiro lugar, os resultados de utilitário de pesquisa da web são analisados com respeito a número de ocorrências ou a probabilidade para cada combinação de caractere alternativa, e a alternativa mais provável é selecionada. De acordo com um exemplo de concretização da presente invenção, um programa executa as etapas de método anteriores se comunicando com a Internet por uma API para um navegador de Internet, provendo as alternativas ortográficas como argumentos de pesquisa, e mede os acertos para as alternativas ortográficas. As alternativas ortográficas como descrito na Figura 7 também podem ser relatadas como um arquivo que pode ser comunicado ao programa de acordo com a presente invenção, como conhecido a uma pessoa qualificada na técnica.

Um exemplo que ilustra a aplicação de uma concretização de acordo com a presente invenção é tomado de uma carta escrita em 1926, e que está armazenada nos Arquivos Nacionais Noruegueses (Riksarkivet). O conteúdo da carta está relacionado a embarque de rena pelo Oceano Atlântico com os navios a vapor Helligolav e Stavangerfjord. Os nomes próprios destes dois navios não podem ser achados em qualquer dicionário de inglês existente. Adicionalmente, neste exemplo de processamento de OCR, o caractere "N" e "H" como ilustrado na Figura 2 é difícil de distinguir. Uma oração da carta de 1926 é ilustrada na Figura 1. Portanto, existem duas alternativas como informado da função de OCR, "Helligolav" e "Nelligolav".

Não existe nenhuma preferência estatística para quaisquer das alternativas em uma estatística de freqüência de letra.

Porém, se forem usadas as duas alternativas "Helligolav" e "Nelligolav" como perguntas em um utilitário de pesquisa da web, há 65 páginas da web contendo a palavra "Helligolav" e nenhuma contendo a palavra sem sentido "Nelligolav", uma verificação clara que a palavra deveria ser reconhecida como "Helligolav". Um dos resultados de pesquisa é um retrato do navio como ilustrado na Figura 3.

De acordo com outro aspecto da presente invenção, conhecimento sobre o conteúdo em um documento a ser reconhecido pode ser usado no processo de confirmação. No exemplo acima, o conhecimento que a carta inclui conteúdo relacionado a navios, animais, etc., pode ser utilizado tal que as perguntas sejam submetidas a sites da Internet incluindo informação relacionada a navios, animais, etc. O retorno de um retrato de uma galeria de quadros incluindo ilustrações de navios é então uma identificação forte sobre o significado da palavra. Um modo de identificar um quadro é identificar a extensão de arquivo como sendo por exemplo ".BMP", ".JPG", etc.

Outro exemplo de uso de uma concretização da presente invenção inclui uma frase do livro popular "Dark Fire" pelo autor C. J. Sansom grafado em uma fonte de letra preta estranha, como descrito na Figura 4. A qualidade da imagem varrida desta oração é de excelente qualidade, e portanto a maioria do texto pode ser decodificado casando símbolos semelhantes e executando uma decifração dos símbolos como uma cifra de substituição monoalfabética, como bem conhecido a uma pessoa qualificada em técnicas usadas em criptoanálise.

As palavras indecifráveis restantes são palavras como o nome próprio "Vaughan", desde que o 'V' é indecifrável porque não há nenhuma outra maiúscula 'V' no texto e a palavra "Vaughan" não é achada em um dicionário. Por estatística de freqüência de letra como conhecido a uma

pessoa qualificada na técnica, as possibilidades das alternativas de confusão de 'V' estão limitadas às letras maiúsculas consoantes 'BCDFGHJKLMNPQRSTVWX'. Os resultados medidos de perguntas de pesquisa da web com estas hipóteses alternativas estão listados na Tabela 1 abaixo.

Tabela 1

Pergunta de Palavra	Resultados de Pergunta (número de páginas da web)
Baughan	629 000 páginas
Caughan	12 300 páginas
Daughan	3 030 páginas
Faughan	32 300 páginas
Gaughan	1240 000 páginas
Haughan	13 800 páginas
Jaughan	45 páginas
Kaughan	199 páginas
Laughan	502 páginas
Maughan	897 000 páginas
Naughan	376 páginas
Paughan	46 páginas
Qaughan	1 page
Raughan	211 páginas
S aughan	63 páginas
Taughan	98 páginas
Vaughan	24 900 000 páginas
Waughan	733 páginas
Xaughan	2 páginas

Embora 'Vaughan' seja mais provável com quase 90% do número total de acertos de pergunta, nenhuma decisão conclusiva pode ser feita diretamente baseado nestes resultados. É possível excluir 'Xaughan e 'Qaughan' como muito improvável por causa do número muito baixo de acertos, mas ainda há uma chance de 10% de uma classificação errônea se o a alternativa 'Vaughan' for selecionada.

Porém se for usada a frase de pesquisa "Vaughan livery" ao invés, foram achadas só 4 páginas contendo a frase com um indício 'V', e nenhuma das outras combinações de caracteres retorna quaisquer acertos de

medição de pergunta. A explicação para estes resultados é que enquanto a família 'Vaughan' faz parte da antiga aristocracia inglesa e conseqüentemente tinha criados em "Vaughan livery", nenhuma das outras famílias Baughan, Caughan, Maughan, etc., tinha criados em sua criadagem como eles não fazem parte da nobreza. Usando conhecimento sobre o conteúdo do texto a ser reconhecido, a palavra mais provável pode ser identificada. Neste exemplo, a palavra "criadagem" é a primeira palavra sucessiva depois da palavra sob investigação. Portanto, apenas combinando esta palavra com todas as outras alternativas possíveis como argumentos de pesquisa, a palavra combinada revela o significado do conteúdo, e conseqüentemente a versão mais provável da palavra sob investigação.

Na Figura 5, é descrito um texto tirado do 'Aenid de Vergil', em que uma das palavras duvidosamente reconhecidas é 'Danae' com a ortografia alternativa 'Danac'. Nenhuma palavra é achada no dicionário. No mesmo texto são reconhecidas certamente as palavras 'Latinus', 'Turnus', 'Rutulian', 'Argos' e 'Long'.

Tabela 2

Palavras	Contagens	*Long*	*Argos*	*Turnus*	*Latinus*	*Rutulian*
Danae	2,510,000	301,000	43,900	584	525	238
Danac	101,000	24,700	130	6	2	0
Contagens		2,270,000,000	11,800,000	1,960,000	807,000	880
Relação	**96%**	**93%**	**99,7%**	**99%**	**99,7%**	**100%**
Co-ocorrência de palavra relativa		**0,01%**	**0,4%**	**0,02%**	**0,06%**	**27%**

Com referência à Tabela 2, a relação de acertos de pesquisa de pergunta da web entre 'Danae' e 'Danac' é 96% a favor de 'Danae', algo que não pode ser visto como conclusivo. Uma possível estratégia é usar pesquisa da web combinando as palavras de pesquisa com as outras palavras reconhecidas certamente. A palavra 'Long' é muito comum e só 0,1 por mil de todos os documentos contendo a palavra 'Long' contêm tanto 'Danae' ou 'Danac', e a relação de acerto é 93%. As palavras 'Argos', 'Turnus' e 'Latinus' estão todas retornando relações de acerto combinadas com 'Danae' e 'Danac'

que favorece 'Danae' (> 99%), mas a co-ocorrência de palavra relativa ainda é pequena. É a palavra menos comum 'Rutulian' que só resulta em 880 acertos sozinha, que conduz a um argumento conclusivo. 'Rutulian' nunca está combinado com 'Danac', mas em 27% dos documentos contendo a palavra 'Rutulian', é também achada a palavra 'Danae', indicando uma forte co-ocorrência de palavra.

A generalização deste princípio é que palavras reconhecidas certamente com baixas contagens de freqüência em perguntas de pesquisa da web que co-ocorrem com um das alternativas de palavra provê resposta mais confiável que palavras reconhecidas certamente com alta freqüência. Geralmente, um aspecto de acordo com a presente invenção é que é possível identificar certamente o que uma palavra não é. Isto é alcançado identificando alternativas que retornam acertos de medição zero da pesquisa na WEB. Geralmente, o número de acertos medidos retornados pode cair dentro de três categorias:

1) O número resultante de acertos medidos está acima de um limiar superior predefinido para uma das alternativas. Então esta alternativa é selecionada.

2) O número de acertos medidos está abaixo de um limiar inferior. Então esta alternativa é descartada.

3) O número de acertos medidos cai entre o limiar superior e inferior. Então a alternativa é investigada adicionalmente.

De acordo com um exemplo de concretização da presente invenção, estas três categorias podem ser usadas como uma medida de confirmação de versão provável de uma palavra sob investigação. De acordo com uma concretização alternativa da presente invenção, o limiar superior e o limiar inferior podem ser variados para cima ou para baixo cooperativamente, ou independentes. Por exemplo, os 100% de acertos totais podem ser divididos em três seções definidas por 10% acima de limiar superior, 10%

abaixo de limiar inferior, que implica que 80% dos acertos caem entre os limiares. De acordo com a concretização alternativa, as gamas podem ser divididas como 5%, 90%, 5%, respectivamente, ou como 10%, 70%, 30%, respectivamente. Qualquer divisão está dentro da extensão da presente invenção.

De acordo com um exemplo de concretização da presente invenção, um método incluindo etapas para confirmar a versão mais provável de uma palavra duvidosamente reconhecida inclui as etapas seguintes:

a) Sempre que um processo de reconhecimento relata um caractere duvidosamente reconhecido, a palavra incluindo este caractere é registrada tal que as alternativas de versão do caractere sejam inseridas na posição do caractere na palavra, por esse meio criando uma lista incluindo alternativas de palavra. Uma função de OCR como conhecido a uma pessoa qualificada na técnica provê tal informação.

b) As palavras na lista são então usadas como perguntas uma a uma em um navegador de Internet como conhecido a uma pessoa qualificada na técnica. Os resultados de pesquisa são medidos e armazenados em uma lista, por exemplo.

c) A próxima etapa é então investigar o resultado na lista de relatório. O processo de seleção de confirmação é baseado na observação que essas pesquisas retornando resultados zero provêem uma certa confirmação sobre o que a palavra não é. Portanto, o processo adicionalmente só investigará aquelas listagens que provêem um resultado de pesquisa diferente de zero. Porém, a interpretação do número de acertos não está só relacionada ao número maior de acertos na Internet, mas em uma taxa de acerto relativa, relativa aos outros acertos. Se a taxa de acerto relativa estiver acima de um limiar superior predefinido para uma alternativa específica, esta alternativa é selecionada como a palavra mais provável.

d) Se a taxa de acerto relativa estiver abaixo do limiar

superior, e a taxa de acerto relativa estiver acima de um limiar de taxa de acerto inferior, investigação adicional é executada. Se a palavra alternativa tiver uma taxa de acerto relativa fora do limiar superior e inferior, a alternativa é tratada como não sendo certamente a palavra.

e) Adicionalmente, investigação da palavra duvidosamente reconhecida inclui as etapas para verificar se a palavra tem uma letra maiúscula, e portanto é um nome próprio provável. Se o processo de reconhecimento retornar outros nomes próprios prováveis, pelo menos dois nomes próprios são usados como uma pergunta de pesquisa combinada. Novamente, a combinação de palavras retornando zero acertos é excluída como sendo candidatas. Os resultados restantes são então testados de acordo com o intervalo de confiança, tanto estando acima de um limiar superior ou abaixo de um limiar inferior, ou como sendo um candidato para investigação adicional quando dentro dos limites de limiar superior e inferior.

f) Se o teste de nome próprio falhar, uma etapa adicional é executar uma combinação de pelo menos uma palavra precedendo e pelo menos uma palavra sucessiva achada no texto relativo à palavra sob investigação. O mesmo teste de confiança é executado.

g) Se o teste de palavra combinada na etapa f) falhar, então pelo menos uma palavra precedendo ou pelo menos uma palavra sucessiva incluindo vários caracteres acima de um limiar predefinido é selecionada para ser combinada com a palavra sob investigação. O teste de confiança é então executado nos resultados informados. Usando só palavras acima de um certo comprimento, palavras pequenas como "a", "o", "e", etc., são evitadas como argumentos de pesquisa.

h) Se o teste de confiança na etapa g) falhar, então uma contagem de freqüência relativa de pelo menos uma palavra precedendo ou pelo menos uma palavras sucessiva é executada, e só aquelas palavras com baixa contagem de freqüência relativa são usadas na etapa g). As medições

para as alternativas ortográficas diferentes são então usadas para avaliar a palavra mais provável, ou são usadas para iniciar medições adicionais de alternativas ortográficas, usando palavra única, combinação de múltiplas palavras, frases e/ou em combinação com curingas como argumentos de pesquisa adicionais que são medidos.

i) Se o teste de confiança falhar na etapa h) e/ou g), então os primeiros caracteres da palavra são usados como entrada a um processo de consulta de dicionário. Quando a combinação de caracteres que retorna um resultado válido do processo de consulta do dicionário é alcançada, esta parte da palavra sob investigação é uma palavra válida que é combinada com as alternativas para a parte restante da palavra. O teste de confiança é então executado novamente.

j) Se quaisquer das etapas c) a i) retornar respostas inconclusas para a palavra sob investigação, o limiar superior e limiar inferior são mudados em etapas cooperativamente várias vezes predefinidas, e as etapas de confirmação c) a i) são repetidas.

k) Se a etapa j) também falhar, seleções aleatórias de limiares superior e inferior são usadas, e as etapas de confirmação c) a i) são repetidas.

l) se o teste de confiança falhar na etapa k), a alternativa tendo a taxa de acerto mais alta da pesquisa resulta na etapa d) é selecionada como a palavra mais provável.

No exemplo de concretização da presente invenção como descrito acima, o caractere duvidosamente reconhecido pode ser dois ou mais caracteres que são difíceis de distinguir. Por exemplo, o caractere "m" pode ser uma combinação de "r" e "n", por exemplo, mas a função de OCR tem problemas em distinguir cada caractere respectivo. Também é uma possibilidade que a função de OCR interprete uma combinação de "r" e "m" distintamente, mas o caractere é de fato "m". Em todas as concretizações da presente invenção, qualquer referência a um caractere duvidosamente

reconhecido pode incluir um ou mais caracteres duvidosamente reconhecidos como ilustrado aqui. Neste contexto, a expressão "alternativa ortográfica" inclui substituição de um caractere duvidosamente reconhecido com uma ou mais substituição possível de um caractere com uma combinação de dois outros caracteres, ou vice-versa.

De acordo com outro aspecto da presente invenção, os valores de limiar usados para determinar a aceitação de uma alternativa ortográfica estão relacionados a medições de possíveis alternativas ortográficas como descrito acima. Porém, o número total de acertos que são medidos em algum senso influenciará o nível atual de limiares que são usados. De acordo com um exemplo de concretização da presente invenção, o nível de aceitação para uma alternativa ortográfica i, denotada como *acceptance(i)* pode ser expresso como:

$$acceptance(i) \Leftrightarrow \frac{\#hits_i}{\sum_{i=1}^{n} \#hits_i} \geq \gamma(\#\text{hits})$$

em que i denota uma das alternativas ortográficas, $\#hits_i$ é o número medido de acertos para alternativa ortográfica i, o denominador é o número medido total de acertos para todas as alternativas ortográficas, e γ (#hits) é um nível de limiar que é uma função do número de acertos.

Em outro exemplo de concretização da presente invenção, o *acceptance(i)* é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que $\max(\#hits_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e os outros parâmetros são como definido acima.

Em um exemplo de concretização da presente invenção, γ é um de dois possíveis valores, um para número muito alto de acertos e outro caso contrário. Em ainda outro exemplo de concretização da presente

invenção, há diferentes γ's para frases, palavras únicas e palavras múltiplas, se a pesquisa incluir curingas, etc., e sempre que uma alternativa ortográfica for medida como uma única palavra, como parte de múltiplas pesquisas de palavra ou como uma frase, os níveis de limiar diferentes são usados respectivamente para verificar a alternativa ortográfica mais provável.

Outra forma do valor de aceitação poderia ser manter a métrica na gama [0,1], um exemplo de limiar pode ser então:

$$acceptance(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \geq \gamma(\#\text{hits}),$$

onde os parâmetros são como definido acima. A definição do limiar também é denotada como rBest(i) usado como argumento em uma função de mérito definida abaixo.

De acordo com outro aspecto da presente invenção, também é possível medir e fazer comparações com níveis de limiar para rejeitar uma alternativa ortográfica, por exemplo usando:

$$rejection(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \leq \kappa(\#\text{hits}),$$

em que os parâmetros são como definido acima, enquanto o nível de limiar inferior como uma função do número de acertos é denotado como κ(#hits).

Em um exemplo de concretização da presente invenção, κ é um de dois possíveis valores, um para número muito alto de acertos e outro caso contrário. Em ainda outro exemplo de concretização da presente invenção, há diferentes κ's para frases, palavras únicas e palavras múltiplas, se a pesquisa incluir curingas, etc., e sempre que uma alternativa ortográfica for medida como uma palavra única, como parte de múltiplas pesquisas de palavra, ou como uma frase, etc., os níveis de limiar diferentes são usados respectivamente para verificar a alternativa ortográfica mais provável.

Como conhecido a uma pessoa qualificada na técnica, programas de OCR também podem informar probabilidades de caractere ou

valores de contagem, denotado valor de CRS, que pode ser usado para designar uma função de mérito que inclui ambos o CRS e #hits das pesquisas de rede. Tais funções de mérito podem ser usadas como valores de aceitação ou valores de rejeição, respectivamente. De acordo com um aspecto da presente invenção, a palavra mais provável é a que maximiza a função de mérito, para palavra i:

$$totscore(i) = aCRS_{word}(i) + b\,\frac{\#hits}{\max(\#hits_j)_{j\neq i}}$$

em que a+b=1, $CRS_{word}(i)$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, $\max(\#hits_j)_{j\neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i. Os fatores de ponderação a e b podem ser usados para regular a importância relativa ou contribuição ao valor de função do valor de CRS e número de acertos, respectivamente.

Uma função de mérito até mesmo mais complicada poderia ser:

$$totscore(i) = a'CRS_{word}(i) + b'(1\text{-}mim(CRSi)) - c'\,\frac{1-\sum_{k=1}^{nchar}\Delta CRS_{i,k}}{nchar} +$$

$$d'\ f(rBest(i)_{phrase}'\ rBest(i)_{\text{single word}},\ rBest(i)_{\text{mult word}})$$

onde o segundo termo é o CRS mínimo para todos os caracteres na palavra, o terceiro termo é a soma da diferença de CRS entre o CRS mais alto para cada caractere e o CRS usando word(i). A função *f* é tanto uma função mínima ou máxima, respectivamente, dos níveis de aceitação diferentes como definido acima relacionado à palavra única i, o nível de aceitação para frases incluindo a palavra i, e pesquisas de múltiplas palavras incluindo a palavra i. Na função a' + b' + c' + d' = 1, e é usada para regular a contribuição de cada elemento, nchar é o número de caracteres em palavras i

De acordo com um aspecto da presente invenção, a expressão "nível de limiar" é para incluir, mas não ser limitada a: um número

selecionado, um número re-normalizado, um nível de aceitação, um valor de contagem total, ou um nível de rejeição.

O método de acordo com a presente invenção como descrito acima pode ser implementado como rotinas de software em um sistema de OCR existente, como conhecido a uma pessoa qualificado na técnica. O único pré-requisito é que a função de reconhecimento relate os caracteres duvidosamente reconhecidos e as palavras incluindo estes caracteres. Adicionalmente, a função de reconhecimento deveria relatar as alternativas para o caractere duvidosamente reconhecido. Adicionalmente, a ordem de etapas de confirmação não tem necessariamente que ser executada como descrito acima, isto é a etapa i) pode ser executada antes da etapa h), como entendido por uma pessoa qualificada na técnica.

De acordo com concretizações da presente invenção, sempre que um argumento de pesquisa é combinado com outras palavras, partes de palavras também podem ser usadas. Adicionalmente, a operação de combinar itens para prover um argumento de pesquisa inclui, mas não está limitada a usar operadores de pesquisa bem conhecidos, por exemplo "casa E dona", em que E é o operador como o argumento de pesquisa, e que é bem conhecido a uma pessoa qualificada na técnica. Adicionalmente, é para ser entendido que também é possível omitir certos tipos de arquivos na pesquisa usando operadores de pesquisa específicos. Por exemplo, prover um "- PDF" depois do argumento de pesquisa omite todos os tipos de arquivos de PDF, que muito freqüentemente incluem imagens de texto varridas. Emitindo um tal comando, o processo de pesquisa evita investigar documentos incluindo os tipos típicos de erros que o processo de pesquisa é visado a corrigir, por esse meio qualificando os documentos usados como base para a verificação como sendo documentos "limpos".

Exemplos adicionais de concretizações da presente invenção incluem um processo de confirmação que primeiro identifica o número de

acertos que palavras precedentes e palavras sucessivas provêem quando usadas como argumentos de pesquisa em um utilitário de pesquisa. Essas palavras sucessivas com baixa taxa de acerto diferente de zero (sob um primeiro limiar), e que incluem um número alto de caracteres (acima de um segundo) limiar, são usadas em combinação com a palavra sob investigação como uma alternativa ortográfica para o processo de confirmação.

De acordo com outro exemplo de concretização da presente invenção, o limiar de confirmação superior e o limiar de confirmação inferior podem ser mudados cooperativamente ou independentes entre si para prover uma afinação dos critérios para categorizar a palavra duvidosamente reconhecida sob investigação. De acordo com este exemplo de concretização, sempre que os limiares são mudados, uma nova pesquisa é iniciada, e o processo é repetido até terminação, tanto quando um resultado excede o limiar superior, ou como um resultado inconcluso, onde a alternativa ortográfica escolhida provendo o número mais alto de acertos é selecionada como a versão mais provável da palavra sob investigação.

De acordo com ainda outro exemplo de çoncretização da presente invenção, um usuário pode selecionar uma gama de sites que o utilitário de pesquisa vai usar ao executar o processo de confirmação. De acordo com esta concretização da presente invenção, não só sites da Internet são selecionáveis, computadores conectados a Intranets, redes de VPR ou redes semelhantes também podem ser selecionadas. De acordo com este exemplo de concretização, toda a autenticação e associações necessárias são executadas na base de informação contida na lista selecionada pelo usuário ao referenciar tais computadores, como conhecido a uma pessoa qualificada na técnica. Também é importante mostrar que as fontes de informação não estão necessariamente limitadas a computador armazenando informação conectado a redes, mas o utilitário de pesquisa de acordo com a presente invenção também pode pesquisar uma unidade de disco rígido conectada localmente ou

remota incluindo informação como esboçado nos princípios da presente invenção. Quer dizer, qualquer sistema de arquivo ou método de montar um sistema de arquivo residindo em computadores locais ou computadores em uma rede é visto como estando dentro da extensão da presente invenção, e como sendo sites pesquisáveis.

Uma pessoa qualificada na técnica pode entender facilmente que o mesmo método e sistemas de acordo com a presente invenção podem ser utilizados em qualquer tipo de sistema de reconhecimento, por exemplo sistemas de reconhecimento de fala. O processo de confirmação pode ser baseado em fonemas, em lugar de caracteres únicos como alternativas de confusão.

Adicionalmente, também é entendido facilmente por uma pessoa qualificada na técnica que etapas semelhantes de acordo com a presente invenção podem ser executadas em um ambiente de verificação ortográfica.

Figura 6 ilustra um exemplo de concretização de um sistema de acordo com a presente invenção como um fluxograma de um programa de computação executando etapas de um método de acordo com a presente invenção provendo uma confirmação de palavra mais provável de uma palavra duvidosamente reconhecida em um sistema de OCR com o qual esta concretização está se comunicando.

Um documento de texto 10 é introduzido a uma máquina de reconhecimento 11 relatando palavras incertas 12 como uma lista de caracteres duvidosamente reconhecidos junto com as palavras, em que estes caracteres foram encontrados. As alternativas ortográficas ou hipóteses são construídas em 13.

As alternativas ortográficas são então usadas como perguntas em pesquisas na WEB em 17.

Alternativamente, as próprias palavras reconhecidas são

registradas em 15. Em 16, um processo adicionando palavras ou frases ou tema/conteúdo ao documento é executado. Junto com as alternativas ortográficas de 18, estas combinações são usadas como argumentos de pesquisa em 17.

A análise em 19 incluindo etapas de confirmação de acordo com a presente invenção é executada nos resultados de pesquisa providos de 17. O processo de seleção em 21 pode usar a medida de confirmação como descrito acima para fazer a seleção atual. Porém, qualquer processo de seleção pode ser implementado de acordo com a presente invenção. Se o processo de seleção for inconcluso, o processo retorna os resultados inconclusos de volta a 16, e o processo continua até que um resultado conclusivo tenha sido alcançado, ou o número de possíveis iterações de estratégias e/ou ajustes de limiar seja exaurido. Então o processo de seleção 21 termina o processo selecionando a alternativa para a palavra sob investigação provendo a medida de confirmação mais alta, e informando esta alternativa de volta à máquina de OCR, que provê um texto completo incluindo todas as palavras duvidosamente reconhecidas confirmadas, substituídas com a alternativa mais provável para cada uma.

De acordo com outro aspecto da presente invenção, um caractere em branco também é visto como sendo um caractere que pode ser um caractere duvidosamente reconhecido. Esta é uma situação em que uma palavra é dividida equivocadamente em duas metades, por exemplo. Está dentro da extensão da presente invenção formar alternativas ortográficas incluindo remover um caractere de uma palavra ou frase.

REIVINDICAÇÕES

1. Método para confirmar palavras duvidosamente reconhecidas relatadas por um processo de Reconhecimento Óptico de Caracteres executando reconhecimento de uma imagem de texto, em que o relatório inclui uma lista de pelo menos um caractere duvidosamente reconhecido junto com alternativas prováveis para este pelo menos um caractere e as palavras em que o caractere foi encontrado, caracterizado pelo fato de que inclui as etapas de:

a) formar alternativas ortográficas para as palavras incluindo o pelo menos um caractere duvidosamente reconhecido substituindo o pelo menos um caractere duvidosamente reconhecido com as alternativas prováveis relatadas para o pelo menos um caractere, um por um e em possíveis combinações em cada palavra encontrada, ou removendo um caractere para formar uma alternativa ortográfica, respectivamente;

b) usar as alternativas ortográficas formadas em a) como argumentos de pesquisa para um utilitário de pesquisa de Internet e medir o número de acertos para resultados de pesquisa para cada alternativa ortográfica de a);

c) comparar os resultados medidos obtidos em b) com um nível de limiar predefinido superior e um nível de limiar predefinido inferior, e sempre que a medição estiver acima do nível de limiar superior, a alternativa ortográfica formada em a) é usada como a correta, e sempre que uma medição cair abaixo do limiar inferior, o resultado é descartado de investigação adicional, e quando a medição cai entre o limiar superior e inferior, investigação adicional é executada selecionando uma estratégia de pesquisa incluindo alternativas ortográficas de a), e então repetir a etapa b) e c).

2. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a etapa a) inclui substituir o pelo menos um caractere

duvidosamente reconhecido com uma combinação de pelo menos dois caracteres ao formar as alternativas ortográficas.

3. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a etapa a) inclui substituir dois ou mais do pelo menos um caractere duvidosamente reconhecido com um único caractere ao formar as alternativas ortográficas.

4. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui identificar se a alternativa ortográfica sob investigação é um nome próprio, e se sim identificar no processo de OCR outras palavras reconhecidas que são nomes próprios, então prover como uma alternativa ortográfica uma combinação da palavra sob investigação junto com pelo menos um outro nome próprio corretamente reconhecido.

5. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

6. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

7. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

8. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

9. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

10. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

11. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

i) obter uma contagem de ocorrência de palavras encontradas na imagem do texto no processo de OCR;

ii) usar a pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um número baixo de ocorrências abaixo de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

12. Método de acordo com a reivindicação 11, caracterizado pelo fato de que a estratégia de pesquisa adicionalmente inclui na etapa ii):

usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um número baixo de ocorrências abaixo de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

13. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

iii) obter uma contagem de ocorrência de palavras encontradas na imagem do texto no processo de OCR;

iv) usar a pelo menos palavra precedente adicionalmente longe

relativa à palavra sob investigação que tem um número alto de ocorrências acima de um primeiro limiar predefinido e que incluem um número alto de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

14. Método de acordo com a reivindicação 13, caracterizado pelo fato de que a estratégia de pesquisa adicionalmente inclui na etapa ii):

usar a pelo menos palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um número alto de ocorrências acima de um primeiro limiar predefinido e que incluem um número alto de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

15. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

v) selecionar palavras precedentes adicionalmente longe relativas à palavra sob investigação uma a uma e listar essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido;

vi) usar as palavras selecionadas listadas em v) como argumentos de pesquisa em um utilitário de pesquisa de Internet e identificar a palavra que provê um número mais baixo de acertos diferentes de zero, e usar essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

16. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

vii) selecionar palavras sucessivas adicionalmente longe relativas à palavra sob investigação uma a uma e listar essas palavras sucessivas que incluem vários caracteres acima de um limiar predefinido;

viii) usar as palavras selecionadas listadas em vii) como um argumento de pesquisa em um utilitário de pesquisa de Internet e identificar a palavra que provê um número mais baixo de acertos diferentes de zero, e usar

essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

17. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a comparação com o limiar superior e a comparação com o limiar inferior é baseada em uma re-normalização dos limiares e número total relatado de acertos.

18. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo cooperativamente, e sempre que uma mudança de limiares é executada, iniciar uma nova pesquisa e processo de confirmação.

19. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo independentemente, e sempre que uma mudança de limiares é executada, iniciar uma nova pesquisa e processo de confirmação.

20. Método de acordo com a reivindicação 1, caracterizado pelo fato de sempre que uma alternativa ortográfica é inconclusa, o resultado ortográfico provendo o número mais alto de acertos re-normalizados relativos é selecionado como a alternativa ortográfica mais provável.

21. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

selecionar caracteres de frente um por um da palavra sob investigação;

combinar estes caracteres em um número crescente de caracteres de frente;

usar cada um dos exemplos de número crescente de caracteres como um argumento para uma consulta de dicionário; e

se o dicionário retornar uma palavra verdadeira da consulta de dicionário, usar esta palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

22. Método de acordo com quaisquer das reivindicações precedentes, caracterizado pelo fato de que o utilitário de pesquisa, como uma alternativa ou além de executar pesquisas na Internet, faz pesquisas em outras fontes de informação não acessíveis pela Internet, mas que são acessíveis por uma Intranet, Rede Privada Virtual, ou redes semelhantes, ou pesquisando diretamente uma unidade de disco rígido conectada incluindo informação.

23. Método de acordo com a reivindicação 22, caracterizado pelo fato de que um usuário pode selecionar de uma lista sites de informação a serem pesquisados durante o processo de confirmação.

24. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits_i}{\sum_{i=1}^{n} \#hits_i} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#hits_i$ é o número medido de acertos para alternativa ortográfica i, o denominador é o número medido total de acertos para todas as alternativas ortográficas, e $\gamma(\#hits)$ é um nível de limiar que é uma função do número de acertos.

25. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#hits_i$ é o número medido de acertos para alternativo ortográfica i, $\max(\#hits_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\gamma(\#hits)$ é um nível de limiar que é uma função do número de acertos.

26. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar inferior é:

$$rejection(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \leq \kappa(\#\text{hits}),$$

em que $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\kappa(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

27. Método de acordo com a reivindicação 1, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = \text{aCRS}_{\text{word}}(\text{i}) + \text{b}\ \frac{\#hits}{\max(\#hits_j)_{j \neq i}}$$

em que a + b = 1, $\text{CRS}_{\text{word}}(\text{i})$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i.

28. Método de acordo com a reivindicação 1, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = \text{a'CRS}_{\text{word}}(\text{i}) + \text{b'(1-mim(CRSi))-c'}\ \frac{1 - \sum_{k=1}^{nchar} \Delta CRS_{i,k}}{nchar} +$$

$$\text{d'}\ \text{f(rBest(i)}_{\text{phrase}}\text{'}\ \text{rBest(i)}_{\text{single word}}, \text{rBest(i)}_{\text{mult word}})$$

em que a' + b' + c' + d' = 1, $\text{CRS}_{\text{word}}(\text{i})$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, o segundo termo é o mínimo CRS para todos os caractere na palavra, o terceiro termo é a soma da diferença de CRS entre o CRS mais alto para cada caractere e o CRS usando palavra (i), *f* é uma função mínima ou máxima dos valores de limiar superior ou limiar inferior como definido de acordo com a reivindicação 25, e *nchar* é o número de caracteres na palavra i

29. Método de acordo com quaisquer das reivindicações 1-28, caracterizado pelo fato de que o sistema de OCR é um sistema de

reconhecimento de fala, e o pelo menos um caractere duvidosamente reconhecido é uma interpretação duvidosa de um fonema.

30. Sistema para confirmar palavras duvidosamente reconhecidas relatadas de uma função de Reconhecimento Óptico de Caracteres (OCR) em um sistema de computador, em que o relatório inclui uma lista para um caractere duvidosamente reconhecido junto com alternativas prováveis para este caractere e as palavras encontradas em um texto incluindo este caractere em um texto sob investigação na função de OCR, caracterizado pelo fato de que inclui:

a) um componente de sistema formando alternativas ortográficas para as palavras incluindo as palavras duvidosamente reconhecidas substituindo o pelo menos um caractere duvidosamente reconhecido com as alternativas prováveis relatadas para o pelo menos um caractere um por um em cada palavra encontrada;

b) um componente de sistema usando as alternativas ortográficas formadas em a) como argumentos de pesquisa para um utilitário de pesquisa de Internet e medindo o número de acertos para resultados de pesquisa para cada alternativa ortográfica;

c) um componente de sistema comparando os resultados medidos obtidos em b) com um nível de limiar predefinido superior e um nível de limiar predefinido inferior, e sempre que a medição estiver acima do nível de limiar superior, a alternativa ortográfica formada em a) é usada como a correta, e sempre que uma medição cair abaixo do limiar inferior, o resultado é descartado de investigação adicional, e quando a medição cai entre o limiar superior e inferior, investigação adicional é executada selecionando uma estratégia de pesquisa com alternativas ortográficas de a), e então executando componente de sistema b) e c).

31. Sistema de acordo com a reivindicação 1, caracterizado pelo fato de que o componente de sistema a) inclui substituir pelo menos um

caractere duvidosamente reconhecido com uma combinação de pelo menos dois caracteres ao formar as alternativas ortográficas.

32. Sistema de acordo com a reivindicação 1, caracterizado pelo fato de que o componente de sistema a) inclui substituir dois ou mais do pelo menos um caractere duvidosamente reconhecido com um único caractere ao formar as alternativas ortográficas.

33. Sistema de acordo com a reivindicação 28, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade identificando se a alternativa ortográfica sob investigação é um nome próprio, e se sim submete uma pergunta ao processo de OCR identificando outras palavras reconhecidas que são nomes próprios, e então combinam pelo menos um dos outros nomes próprios reconhecidos corretamente com o nome próprio sob investigação como a alternativa ortográfica.

34. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema de pesquisa b) inclui uma unidade usando pelo menos uma palavra precedente relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

35. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade usando pelo menos uma palavra sucessiva relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

36. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade usando pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

37. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade usando pelo

menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

38. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade usando pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

39. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema b) inclui uma unidade usando pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

40. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema de pesquisa b) inclui uma unidade que:

ix) obtém uma contagem de ocorrência de palavras encontradas na imagem do texto do processo de OCR, e armazena os números de ocorrência;

x) seleciona pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um baixo número de ocorrência de v) sob um limiar predefinido e combina esta palavra com a palavra sob investigação como a alternativa ortográfica.

41. Método de acordo com a reivindicação 28, caracterizado pelo fato de que o componente de sistema b) adicionalmente inclui uma unidade que:

seleciona pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um baixo número de ocorrência de ix) sob um limiar predefinido e combina esta palavra com a

palavra sob investigação como a alternativa ortográfica.

42. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema de pesquisa b) inclui uma unidade que:

xi) obtém uma contagem de ocorrência de palavras encontradas na imagem do texto do processo de OCR, e armazena os números de ocorrência;

xii) seleciona pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um alto número de ocorrências acima de um primeiro limiar predefinido e que inclui um alto número de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

43. Sistema de acordo com a reivindicação 42, caracterizado pelo fato de que o componente de sistema de pesquisa b) adicionalmente inclui uma unidade que:

seleciona pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um alto número de ocorrências acima de um primeiro limiar predefinido e que inclui um alto número de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

44. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o componente de sistema de pesquisa b) inclui uma unidade que:

xiii) seleciona palavras precedentes adicionalmente longe relativas à palavra sob investigação uma a uma e armazena essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido,

xiv) usa as palavras armazenadas em xiii) como argumentos de pesquisa em um utilitário de pesquisa de Internet, identifica a palavra que provê um número mais baixo de acertos diferentes de zero, e usa essa palavra

em combinação com a palavra sob investigação como a alternativa ortográfica.

45. Sistema de acordo com a reivindicação 28, caracterizado pelo fato de que o componente de sistema de pesquisa b) inclui uma unidade que:

xv) seleciona palavras sucessivas adicionalmente longe relativas à palavra sob investigação uma a uma e armazena essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido;

xvi) usa as palavras armazenadas de xv) como argumentos de pesquisa em um utilitário de pesquisa de Internet, identifica a palavra que provê um número mais baixo de acertos diferentes de zero, e usa essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

46. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que a comparação com o limiar superior e a comparação com o limiar inferior é baseada em uma re-normalização dos limiares e números totais medidos de acertos.

47. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo cooperativamente.

48. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo independentemente.

49. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de sempre que uma alternativa ortográfica está inconclusa, o resultado ortográfico provendo o número mais alto de acertos relativos (re-normalizados) é selecionado como a alternativa ortográfica mais provável.

50. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o utilitário de pesquisa, como uma alternativa ou além de

executar pesquisas na Internet, faz pesquisas em outras fontes de informação não acessíveis pela Internet, mas que são acessíveis por uma Intranet, VPR, ou redes semelhantes, ou diretamente pesquisando uma unidade de disco rígido conectada incluindo informação.

51. Sistema de acordo com a reivindicação 50, caracterizado pelo fato de que um usuário pode selecionar de uma lista uma gama de sites de informação a serem pesquisados durante o processo de confirmação pelo utilitário de pesquisa.

52. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota um das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, o denominador é o número medido total de acertos para todas as alternativas ortográfica, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

53. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

54. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que o limiar inferior é definido como:

$$rejection(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \leq \kappa(\#\text{hits}),$$

em que $\#\text{hits}_i$ é o número medido de acertos para alternativo

ortográfica i, $\max(\#hits_j)_{j\neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\kappa(\#hits)$ é um nível de limiar que é uma função do número de acertos.

55. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = a\mathrm{CRS}_{word}(i) + b\,\frac{\#hits}{\max(\#hits_j)_{j\neq i}}$$

é em que a + b = 1, $\mathrm{CRS}_{word}(i)$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, $\max(\#hits_j)_{j\neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i.

56. Sistema de acordo com a reivindicação 30, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = a'\mathrm{CRS}_{word}(i) + b'(1\text{-}mim(CRSi))\text{-}c'\,\frac{1-\sum_{k=1}^{nchar}\Delta CRS_{i,k}}{nchar} +$$

$$d'\ f(\mathrm{rBest(i)}_{phrase}'\ \mathrm{rBest(i)}_{\text{single word}},\ \mathrm{rBest(i)}_{\text{mult word}})$$

em que a' + b' + c'+ d' = 1, $\mathrm{CRS}_{word}(i)$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, o segundo termo é o CRS mínimo para todos os caracteres na palavra, o terceiro termo é a soma da diferença de CRS entre o CRS mais alto para cada caractere e o CRS usando palavra (i), *f* é uma função mínima ou máxima dos valores de limiar superior ou limiar inferior como definido de acordo com a reivindicação 54, e *nchar* é o número de caracteres na palavra i.

57. Sistema de acordo com reivindicações 30-56, caracterizado pelo fato de que o sistema de OCR é um sistema de reconhecimento de fala, e o pelo menos um caractere duvidosamente reconhecido é uma interpretação duvidosa de um fonema.

Figura 1

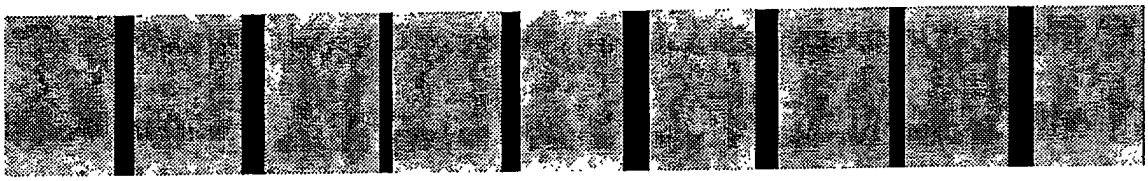


Figura 2

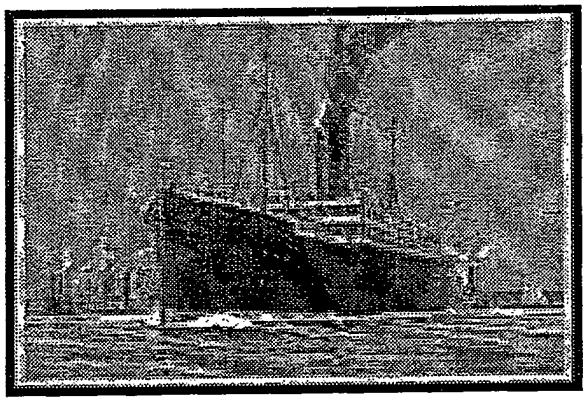
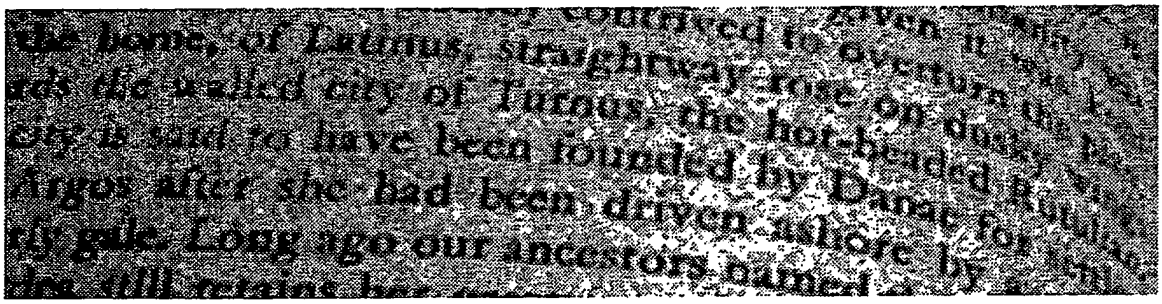


Figura 3

The Vatican Library

Figura 4



the home of Latinus, straightway rose on overturn the
eds the walled city of Turnus, the hot-headed Rutul
city is said to have been founded by Danae for
Argos after she had been driven ashore by
tly gale. Long ago our ancestors named
dra still retains her

Figura 5

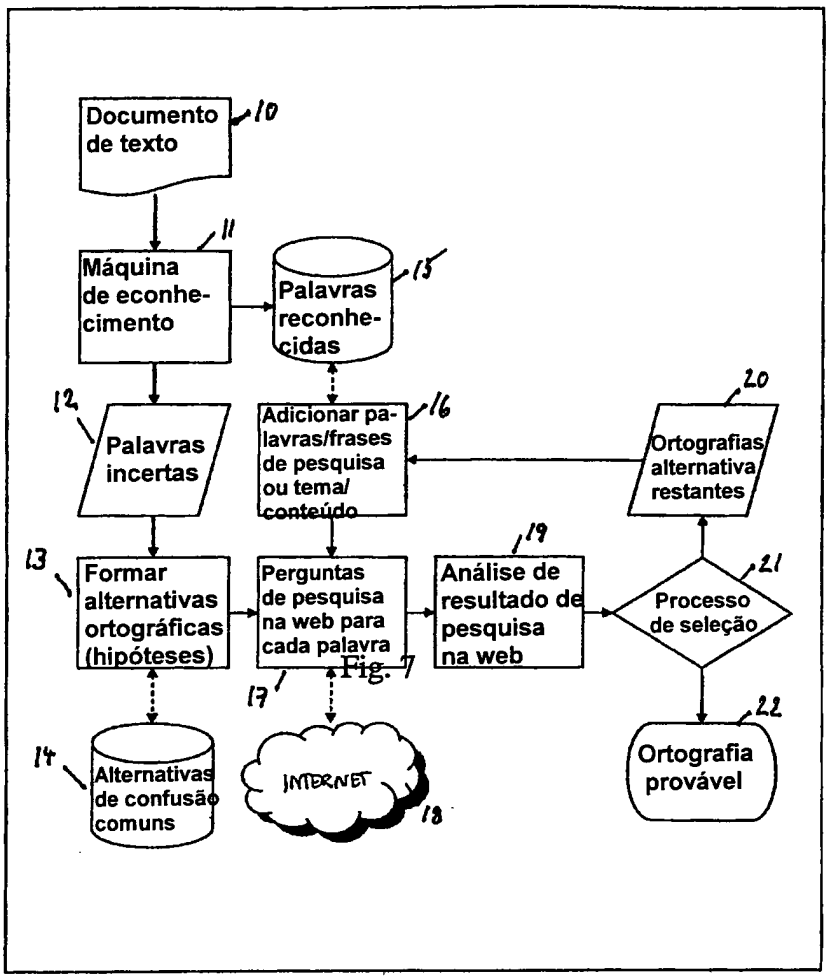


Fig. 7

Figura 6

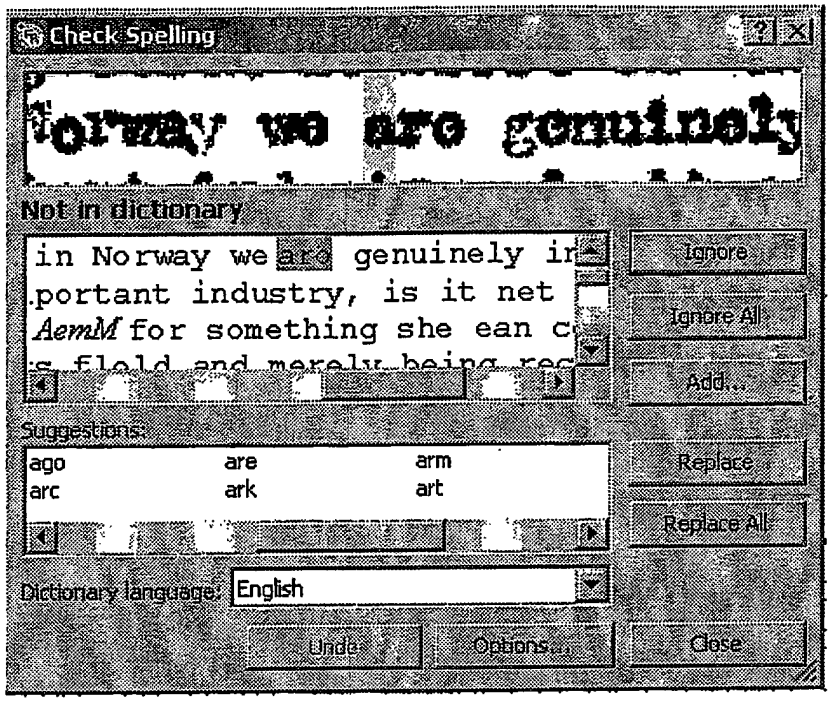

Check Spelling
Not in dictionary:
in Norway we are genuinely in
portant industry, is it net
AemM for something she ean c
Ignore
Ignore All
Add...
Suggestions:
ago
are
arm
arc
ark
art
Replace
Replace All
Dictionary language:
English
Undo
Options...
Close

Figura 7

RESUMO

"MÉTODO E SISTEMA PARA RESOLVER DADOS DE SAÍDA CONTRADITÓRIOS DE UM SISTEMA DE RECONHECIMENTO ÓPTICO DE CARACTERES"

A presente invenção provê um método e sistema para confirmar palavras duvidosamente reconhecidas como relatadas por um processo de Reconhecimento Óptico de Caracteres usando alternativas ortográficas como argumentos de pesquisa para um utilitário de pesquisa de Internet. O número medido de acertos para cada alternativa ortográfica é usado para prover uma medida de confirmação para a alternativa ortográfica mais provável. Sempre que a medida de confirmação é inconclusa, uma pluralidade de estratégias de pesquisa é usada para alcançar um resultado medido incluindo acertos zero, exceto para uma alternativa ortográfica que é usada como a alternativa correta.

A requerente apresenta novas vias das reivindicações para melhor esclarecer e definir o presente pedido.

REIVINDICAÇÕES

1. Método para resolver dados de saída contraditórios de um sistema de Reconhecimento Óptico de Caracteres (OCR), em que os dados de saída compreendem pelo menos uma palavra com pelo menos um caractere duvidosamente reconhecido, em que o pelo menos um caractere duvidosamente reconhecido é relatado nos dados de saída junto com alternativas prováveis para o pelo menos um caractere duvidosamente reconhecido, e as palavras nas quais este pelo menos um caractere duvidosamente reconhecido foi encontrado em uma imagem de um texto sendo processado pelo sistema de OCR, caracterizado pelo fato de que compreende as etapas de:

usar um utilitário de pesquisa de Internet com argumentos de pesquisa estabelecidos de acordo com uma estratégia de pesquisa compreendendo:

a) fornecer argumentos de pesquisa iniciais formando alternativas ortográficas para as palavras incluindo o pelo menos um caractere duvidosamente reconhecido substituindo o pelo menos um caractere duvidosamente reconhecido com as alternativas prováveis relatadas para o pelo menos um caractere, um por um e em possíveis combinações em cada palavra encontrada, ou removendo um caractere de modo a formar uma pluralidade de alternativas ortográficas, e então medir e gravar número de acertos para resultados de pesquisa de cada alternativa ortográfica respectiva que foi formada desta maneira,

b) comparar o número medido de acertos para cada uma das alternativas ortográficas com um nível de limiar predefinido superior relativo e um nível de limiar predefinido inferior relativo, em que cada uma das comparações respectivas da pluralidade de medições caem em um dos três possíveis resultados:

i) se a medição de uma alternativa ortográfica estiver acima do

nível de limiar predefinido superior relativo predefinido, a alternativa ortográfica correspondente para esta medição é a alternativa ortográfica correta para a palavra, e termina a pesquisa da Internet,

ii) se a medição de uma alternativa ortográfica estiver abaixo do nível de limiar predefinido inferior relativo predefinido, a alternativa ortográfica correspondente para esta medição é considerada não existir, e a palavra com esta alternativa ortográfica é descartada de investigações adicionais, e continua com outras alternativas ortográficas que foram formadas como argumentos de pesquisa para o utilitário de pesquisa de Internet,

iii) se a medição de uma alternativa ortográfica cair entre o nível de limiar relativo superior e o nível de limiar relativo inferior, sair do utilitário de pesquisa de Internet e modificar a estratégia de pesquisa fornecendo argumentos de pesquisa adicionais como uma combinação de números das alternativas ortográficas remanescentes e outras palavras encontradas no documento, outras alternativas de caractere para o pelo menos um caractere duvidosamente reconhecido, frases, adaptar o nível de limiar relativo superior, adaptar o nível de limiar relativo superior, e/ou outra informação relacionada aos dados de saída do sistema de OCR, antes de continuar usando a estratégia de pesquisa fornecendo medições e comparações adicionais para resolver os dados de saída contraditórios,

c) continuar processar etapa b) um número de tempos predefinidos ou até que exista somente uma alternativa ortográfica sobrando, o que sempre ocorrer primeiro, fornecer uma iteração entre uma pluralidade de diferentes argumentos de pesquisa usados na estratégia de pesquisa antes de terminar a etapa b), e usar a alternativa ortográfica remanescente tendo a mais alta medição acima do nível de limiar superior relativo como a alternativa ortográfica correta.

2. Método de acordo com a reivindicação 1, caracterizado pelo

fato de que a estratégia de pesquisa inclui substituir o pelo menos um caractere duvidosamente reconhecido com uma combinação de pelo menos dois caracteres ao formar as alternativas ortográficas.

3. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui substituir dois ou mais do pelo menos um caractere duvidosamente reconhecido com um único caractere ao formar as alternativas ortográficas.

4. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui identificar se a alternativa ortográfica sob investigação é um nome próprio, e se sim identificar no processo de OCR outras palavras reconhecidas que são nomes próprios, então prover como uma alternativa ortográfica uma combinação da palavra sob investigação junto com pelo menos um outro nome próprio corretamente reconhecido.

5. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

6. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

7. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

8. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação em

combinação com a palavra sob investigação como a alternativa ortográfica.

9. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

10. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

11. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

i) obter uma contagem de ocorrência de palavras encontradas na imagem do texto no processo de OCR;

ii) usar a pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um número baixo de ocorrências abaixo de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

12. Método de acordo com a reivindicação 11, caracterizado pelo fato de que a estratégia de pesquisa adicionalmente inclui na etapa ii):

usar pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um número baixo de ocorrências abaixo de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

13. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

i) obter uma contagem de ocorrência de palavras encontradas na imagem do texto no processo de OCR;

ii) usar a pelo menos palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um número alto de ocorrências acima de um primeiro limiar predefinido e que incluem um número alto de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

14. Método de acordo com a reivindicação 13, caracterizado pelo fato de que a estratégia de pesquisa adicionalmente inclui na etapa ii):

usar a pelo menos palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um número alto de ocorrências acima de um primeiro limiar predefinido e que incluem um número alto de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

15. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

i) selecionar palavras precedentes adicionalmente longe relativas à palavra sob investigação uma a uma e listar essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido;

ii) usar as palavras selecionadas listadas em i) como argumentos de pesquisa em um utilitário de pesquisa de Internet e identificar a palavra que provê um número mais baixo de acertos diferentes de zero, e usar essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

16. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

i) selecionar palavras sucessivas adicionalmente longe relativas à palavra sob investigação uma a uma e listar essas palavras sucessivas que incluem vários caracteres acima de um limiar predefinido;

ii) usar as palavras selecionadas listadas em i) como um argumento de pesquisa em um utilitário de pesquisa de Internet e identificar a

palavra que provê um número mais baixo de acertos diferentes de zero, e usar essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

17. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a comparação com o limiar superior e a comparação com o limiar inferior é baseada em uma re-normalização dos limiares e número total relatado de acertos.

18. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo cooperativamente, e sempre que uma mudança de limiares é executada, iniciar uma nova pesquisa e processo de confirmação.

19. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo independentemente, e sempre que uma mudança de limiares é executada, iniciar uma nova pesquisa e processo de confirmação.

20. Método de acordo com a reivindicação 1, caracterizado pelo fato de que a estratégia de pesquisa inclui as etapas de:

selecionar caracteres de frente um por um da palavra sob investigação;

combinar estes caracteres em um número crescente de caracteres de frente;

usar cada um dos exemplos de número crescente de caracteres como um argumento para uma consulta de dicionário; e

se o dicionário retornar uma palavra verdadeira da consulta de dicionário, usar esta palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

21. Método de acordo com quaisquer das reivindicações precedentes, caracterizado pelo fato de que o utilitário de pesquisa, como uma alternativa ou além de executar pesquisas na Internet, faz pesquisas em outras

fontes de informação não acessíveis pela Internet, mas que são acessíveis por uma Intranet, Rede Privada Virtual, ou redes semelhantes, ou pesquisando diretamente uma unidade de disco rígido conectada incluindo informação.

22. Método de acordo com a reivindicação 21, caracterizado pelo fato de que um usuário pode selecionar de uma lista que possui sites de informação a serem pesquisados durante o processo de confirmação.

23. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits_i}{\sum_{i=1}^{n} \#hits_i} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, o denominador é o número medido total de acertos para todas as alternativas ortográficas, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

24. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativo ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

25. Método de acordo com a reivindicação 1, caracterizado pelo fato de que o limiar inferior é:

$$rejection(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \leq \kappa(\#\text{hits}),$$

em que $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as

alternativas ortográficas não incluindo a alternativa ortográfica para i, e κ(#hits) é um nível de limiar que é uma função do número de acertos.

26. Método de acordo com a reivindicação 1, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = aCRS_{word}(i) + b\,\frac{\#hits}{\max(\#hits_j)_{j\neq i}}$$

em que a + b = 1, $CRS_{word}(i)$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, $\max(\#hits_j)_{j\neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i.

27. Método de acordo com a reivindicação 1, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = a'CRS_{word}(i) + b'(1\text{-}mim(CRSi))\text{-}c'\,\frac{1-\sum_{k=1}^{nchar}\Delta CRS_{i,k}}{nchar} +$$

$$d'\ f(rBest(i)_{phrase}'\ rBest(i)_{\ single\ word},\ rBest(i)_{mult\ word})$$

em que a' + b' + c' + d' = 1, $CRS_{word}(i)$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, o segundo termo é o mínimo CRS para todos os caractere na palavra, o terceiro termo é a soma da diferença de CRS entre o CRS mais alto para cada caractere e o CRS usando palavra (i), *f* é uma função mínima ou máxima dos valores de limiar superior ou limiar inferior como definido de acordo com a reivindicação 25, e *nchar* é o número de caracteres na palavra i

28. Método de acordo com quaisquer das reivindicações 1-27, caracterizado pelo fato de que o sistema de OCR é um sistema de reconhecimento de fala, e o pelo menos um caractere duvidosamente reconhecido é uma interpretação duvidosa de um fonema.

29. Sistema para resolver dados de saída contraditórios de um

sistema de Reconhecimento Óptico de Caracteres (OCR), em que os dados de saída compreendem pelo menos uma palavra com pelo menos um caractere duvidosamente reconhecido, em que o pelo menos um caractere duvidosamente reconhecido é relatado nos dados de saída junto com alternativas prováveis para o pelo menos um caractere duvidosamente reconhecido, e as palavras nas quais este pelo menos um caractere duvidosamente reconhecido foi encontrado em uma imagem de um texto sendo processado pelo sistema de OCR, caracterizado pelo fato de que compreende:

um componente de sistema usando um utilitário de pesquisa de Internet com argumentos de pesquisa estabelecidos de acordo com uma estratégia de pesquisa compreendendo:

a) o componente de sistema fornece argumentos de pesquisa iniciais formando alternativas ortográficas para as palavras incluindo o pelo menos um caractere duvidosamente reconhecido substituindo o pelo menos um caractere duvidosamente reconhecido com as alternativas prováveis relatadas para o pelo menos um caractere, um por um e em possíveis combinações em cada palavra encontrada, ou removendo um caractere de modo a formar uma pluralidade de alternativas ortográficas, e então medir e gravar número de acertos para resultados de pesquisa de cada alternativa ortográfica respectiva que foi formada desta maneira,

b) o componente de sistema compara o número medido de acertos para cada uma das alternativas ortográficas com um nível de limiar predefinido superior relativo e um nível de limiar predefinido inferior relativo, em que cada uma das comparações respectivas da pluralidade de medições caem em um dos três possíveis resultados:

i) se a medição de uma alternativa ortográfica estiver acima do nível de limiar predefinido superior relativo predefinido, a alternativa ortográfica correspondente para esta medição é a alternativa ortográfica

correta para a palavra, e termina a pesquisa da Internet,

ii) se a medição de uma alternativa ortográfica estiver abaixo do nível de limiar predefinido inferior relativo predefinido, a alternativa ortográfica correspondente para esta medição é considerada não existir, e a palavra com esta alternativa ortográfica é descartada de investigações adicionais, e continua com outras alternativas ortográficas que foram formadas como argumentos de pesquisa para o utilitário de pesquisa de Internet,

iii) se a medição de uma alternativa ortográfica cair entre o nível de limiar relativo superior e o nível de limiar relativo inferior, sair do utilitário de pesquisa de Internet e modificar a estratégia de pesquisa fornecendo argumentos de pesquisa adicionais como uma combinação de números das alternativas ortográficas remanescentes e outras palavras encontradas no documento, outras alternativas de caractere para o pelo menos um caractere duvidosamente reconhecido, frases, adaptar o nível de limiar relativo superior, adaptar o nível de limiar relativo superior, e/ou outra informação relacionada aos dados de saída do sistema de OCR, antes de continuar usando a estratégia de pesquisa fornecendo medições e comparações adicionais para resolver os dados de saída contraditórios,

c) o componente de sistema é processar etapa b) um número de tempos predefinidos ou até que exista somente uma alternativa ortográfica sobrando, o que sempre ocorrer primeiro, fornecer uma iteração entre uma pluralidade de diferentes argumentos de pesquisa usados na estratégia de pesquisa antes de terminar a etapa b), e usar a alternativa ortográfica remanescente tendo a mais alta medição acima do nível de limiar superior relativo como a alternativa ortográfica correta.

30. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui substituir pelo menos um caractere duvidosamente reconhecido com uma combinação de pelo menos

dois caracteres ao formar as alternativas ortográficas.

31. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui substituir dois ou mais do pelo menos um caractere duvidosamente reconhecido com um único caractere ao formar as alternativas ortográficas.

32. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade identificando se a alternativa ortográfica sob investigação é um nome próprio, e se sim submete uma pergunta ao processo de OCR identificando outras palavras reconhecidas que são nomes próprios, e então combinam pelo menos um dos outros nomes próprios reconhecidos corretamente com o nome próprio sob investigação como a alternativa ortográfica.

33. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema de pesquisa inclui uma unidade usando pelo menos uma palavra precedente relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

34. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade usando pelo menos uma palavra sucessiva relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

35. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade usando pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

36. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade usando pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob

investigação em combinação com a palavra sob investigação como a alternativa ortográfica.

37. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade usando pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

38. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade usando pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que inclui vários caracteres acima de um limiar predefinido em combinação com a palavra sob investigação como a alternativa ortográfica.

39. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade que:

i) obtém uma contagem de ocorrência de palavras encontradas na imagem do texto do processo de OCR, e armazena os números de ocorrência;

ii) seleciona pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um baixo número de ocorrência de i) sob um limiar predefinido e combina esta palavra com a palavra sob investigação como a alternativa ortográfica.

40. Sistema de acordo com a reivindicação 39, caracterizado pelo fato de que o componente de sistema adicionalmente inclui uma unidade que:

seleciona pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um baixo número de ocorrência sob um limiar predefinido e combina esta palavra com a palavra sob investigação como a alternativa ortográfica.

41. Sistema de acordo com a reivindicação 29, caracterizado

pelo fato de que o componente de sistema inclui uma unidade que:

i) obtém uma contagem de ocorrência de palavras encontradas na imagem do texto do processo de OCR, e armazena os números de ocorrência;

ii) seleciona pelo menos uma palavra precedente adicionalmente longe relativa à palavra sob investigação que tem um alto número de ocorrências acima de um primeiro limiar predefinido e que inclui um alto número de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

42. Sistema de acordo com a reivindicação 41, caracterizado pelo fato de que o componente de sistema adicionalmente inclui uma unidade que:

seleciona pelo menos uma palavra sucessiva adicionalmente longe relativa à palavra sob investigação que tem um alto número de ocorrências acima de um primeiro limiar predefinido e que inclui um alto número de caracteres na palavra acima de um segundo limiar em combinação com a palavra sob investigação como a alternativa ortográfica.

43. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade que:

i) seleciona palavras precedentes adicionalmente longe relativas à palavra sob investigação uma a uma e armazena essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido,

ii) usa as palavras armazenadas em i) como argumentos de pesquisa em um utilitário de pesquisa de Internet, identifica a palavra que provê um número mais baixo de acertos diferentes de zero, e usa essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

44. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema inclui uma unidade que:

i) seleciona palavras sucessivas adicionalmente longe relativas à palavra sob investigação uma a uma e armazena essas palavras precedentes que incluem vários caracteres acima de um limiar predefinido;

ii) usa as palavras armazenadas de i) como argumentos de pesquisa em um utilitário de pesquisa de Internet, identifica a palavra que provê um número mais baixo de acertos diferentes de zero, e usa essa palavra em combinação com a palavra sob investigação como a alternativa ortográfica.

45. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que a função fornece comparação com o limiar superior e a comparação com o limiar inferior é baseada em uma re-normalização dos limiares e números totais medidos de acertos.

46. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo cooperativamente.

47. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o limiar superior e inferior é mudado incrementalmente para cima e para baixo independentemente.

48. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de sempre que uma alternativa ortográfica está inconclusa, o resultado ortográfico provendo o número mais alto de acertos relativos (re-normalizados) é selecionado como a alternativa ortográfica mais provável.

49. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o componente de sistema, como uma alternativa ou além de executar pesquisas na Internet, faz pesquisas em outras fontes de informação não acessíveis pela Internet, mas que são acessíveis por uma Intranet, VPR, ou redes semelhantes, ou diretamente pesquisando uma unidade de disco rígido conectada incluindo informação.

50. Sistema de acordo com a reivindicação 49, caracterizado

pelo fato de que um usuário pode selecionar de uma lista uma gama de sites de informação a serem pesquisados durante o processo de confirmação pelo componente de sistema.

51. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota um das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, o denominador é o número medido total de acertos para todas as alternativas ortográfica, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

52. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o limiar superior é definido como:

$$acceptance(i) \Leftrightarrow \frac{\#hits}{\max(\#hits_j)_{j \neq i}} \geq \gamma(\#\text{hits}),$$

em que i denota uma das alternativas ortográficas, $\#\text{hits}_i$ é o número medido de acertos para alternativa ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\gamma(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

53. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que o limiar inferior é definido como:

$$rejection(i) \Leftrightarrow \frac{\#hits_i}{\#hits_i + \max(\#hits_j)_{j \neq i}} \equiv \text{rBest(i)} \leq \kappa(\#\text{hits}),$$

em que $\#\text{hits}_i$ é o número medido de acertos para alternativo ortográfica i, $\max(\#\text{hits}_j)_{j \neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i, e $\kappa(\#\text{hits})$ é um nível de limiar que é uma função do número de acertos.

54. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para

o número de acertos como:

$$totscore(i) = a\mathrm{CRS}_{\mathrm{word}}(\mathrm{i}) + \mathrm{b}\ \frac{\#hits}{\max(\#hits_j)_{j\neq i}}$$

é em que a + b = 1, $\mathrm{CRS}_{\mathrm{word}}(\mathrm{i})$ é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, $\max(\#\mathrm{hits_j})_{j\neq i}$ é o número medido total de acertos para todas as alternativas ortográficas não incluindo a alternativa ortográfica para i.

55. Sistema de acordo com a reivindicação 29, caracterizado pelo fato de que uma função de mérito é usada para definir uma medição para o número de acertos como:

$$totscore(i) = \mathrm{a'CRS}_{\mathrm{word}}(\mathrm{i}) + \mathrm{b'(1\text{-}mim(CRSi))\text{-}c'}\ \frac{1-\sum_{k=1}^{nchar}\Delta CRS_{i,k}}{nchar} +$$

$$\mathrm{d'\ f(rBest(i)_{phrase}'\ rBest(i)_{\ single\ word},\ rBest(i)_{mult\ word})}$$

em que a' + b' + c'+ d' = 1, $\mathrm{CRS}_{\mathrm{word}}$ (i) é um valor de contagem de caractere do processo de OCR relacionado à alternativa ortográfica i, o segundo termo é o CRS mínimo para todos os caracteres na palavra, o terceiro termo é a soma da diferença de CRS entre o CRS mais alto para cada caractere e o CRS usando palavra (i), *f* é uma função mínima ou máxima dos valores de limiar superior ou limiar inferior como definido de acordo com a reivindicação 53, e *nchar* é o número de caracteres na palavra i.

56. Sistema de acordo com reivindicações 29-55, caracterizado pelo fato de que o sistema de OCR é um sistema de reconhecimento de fala, e o pelo menos um caractere duvidosamente reconhecido é uma interpretação duvidosa de um fonema.