Uso de marcadores estilísticos para a busca na Internet em português

Rachel Aires
rachelaires@gmail.com

Este documento resume a tese de doutorado ***Uso de Marcadores Estilísticos para a busca na Web em português***[1] [6] desenvolvida por dois anos no pólo de Oslo da Linguateca e por dois anos no NILC (Núcleo Interinstitucional de Lingüística Computacional)[2]. O trabalho de doutorado teve como principal objetivo pesquisar uma maneira de minimizar consideravelmente um dos principais problemas dos usuários de sistemas de busca na Internet, que é ter que lidar com um grande volume de documentos irrelevantes para ter acesso à informação procurada. Para que um documento seja relevante, não basta que ele trate do assunto procurado, é necessário ainda que dê o enfoque desejado pelo usuário. O enfoque pode ser determinado por características como, por exemplo, formalidade, objetividade e o fato de o texto ser detalhado, tratando apenas de um assunto e não de vários.

A solução explorada para auxiliar o usuário a interpretar qual o enfoque sobre o assunto procurado é dado por um determinado texto foi classificá-los em **gêneros**, **tipos de textos**, **necessidades de busca** e **necessidades personalizadas**. Para gerar os métodos de classificação, foram utilizados algoritmos de aprendizado de máquina, corpora compilados com textos em português e marcadores estilísticos.

Para a classificação em **gêneros** e para a classificação em **tipos textuais**, foram utilizados os gêneros e tipos textuais do corpus Lácio-Ref [7]. O Lácio-Ref é um corpus aberto e de referência do português contemporâneo do Projeto Lácio-Web, composto de textos em português brasileiro, tendo como característica serem escritos respeitando a norma culta. A taxonomia de gêneros do Lácio-Ref é composta por textos científicos, de referência, informativos, jurídicos, prosa, poesia, drama, instrucionais e técnico-administrativos. Entretanto a versão utilizada do corpus não contém textos do gênero de referência ou do gênero técnico-administrativo. Nos experimentos com classificação em gêneros, os gêneros poesia, prosa e drama foram reunidos em um único supergênero Literário. São 29 os tipos textuais efetivamente presentes na versão utilizada do Lácio-Ref: declaração, petição, reportagem, artigo, decreto, medida provisória, poema, resenha, edital, portaria, resolução, editorial, monografia, resumo, carta, provimento, sentença, circular, entrevista, notícia, receita, súmula, lei, ofício, regimento, crônica, livro-texto, parecer, relatório.

O esquema de classificação em **necessidades de busca** é resultado de uma análise qualitativa dos *logs* de novembro de 1999 e de julho de 2002 da máquina de busca TodoBr[3]. Os sete tipos de necessidades tratados são: encontrar (1) Páginas que definam alguma coisa ou ensinem como e/ou porque algo acontece. Por exemplo: o que é a aurora boreal. Para esta necessidade, os melhores resultados seriam dicionários e enciclopédias, livros didáticos, artigos técnicos e relatórios e textos do gênero informativo; (2) Páginas que ensinem como fazer algo ou como algo é feito. Por exemplo: instruções de como instalar Linux em seu computador, receita

[1] O doutorado foi financiado pela Fundação para Computação Científica Nacional (FCCN), através da Fundação para a Ciência e Tecnologia e co-financiado pelo POSI (POSI/PLP/43931/2001).
[2] www.nilc.icmc.usp.br
[3] Máquina de busca do domínio .br que foi incorporada ao Google em 2005, como Google Brasil: http://www.todobr.com.br/

de um bolo. Resultados típicos seriam textos do gênero instrucional, tais como manuais, livros didáticos, receitas e também alguns artigos técnicos e relatórios; (3) Páginas que forneçam uma apresentação (ou apanhado ou panorama) sobre um determinado assunto. Por exemplo, um panorama sobre a literatura americana no século XX. Nesse caso, os melhores textos seriam dos gêneros instrucionais, informativo e científico, por exemplo, reportagens; (4) Páginas com notícias. Por exemplo: uma notícia sobre um atentado. As melhores respostas seriam textos do gênero informativo, como, por exemplo, notícias em jornais e revistas; (5) Páginas que forneçam informações sobre uma pessoa, ou empresa, ou instituição, ou organização. Por exemplo: páginas pessoais, páginas com informações para contato (com currículo, telefone, endereço). Respostas típicas seriam páginas pessoais e institucionais; (6) Uma página específica que o usuário quer visitar, mas não se lembra da URL. Nesse caso, os resultados poderiam ser de qualquer tipo textual ou gênero; (7) Páginas que forneçam algum serviço online. Por exemplo: lojas virtuais, serviço dos correios para acompanhamento de envio de encomendas. As melhores respostas, nesse caso, seriam textos comerciais (empresas ou indivíduos oferecendo produtos e serviços).

Para gerar um classificador que considerasse as necessidades acima um corpus foi criado inicialmente com 511 textos coletados da Internet obedecendo a dois critérios: (i) as páginas deveriam ser escritas em português do Brasil, para que variações lexicais, morfológicas e sintáticas entre as diversas variantes não interferissem no treinamento dos classificadores e (ii) as páginas selecionadas deveriam ser de diversas fontes e assuntos, já que textos de uma mesma fonte ou área podem ter estilo próprio (por exemplo, textos da Folha de São Paulo e textos médicos), sendo que o que pretendíamos investigar eram os marcadores de estilo relacionados ao propósito do texto. Nessa primeira versão do corpus, não foi considerado o fato de que um mesmo texto pode atender a mais de uma necessidade. Por isso, todos os textos foram revistos e novos textos foram incluídos para cada uma das combinações dos tipos de necessidades tratados encontradas. O corpus criado contém 1703 textos extraídos da Internet brasileira classificados conforme as necessidades de usuários a que satisfazem.

Investigou-se também a possibilidade de permitir a criação de esquemas de categorização pelo próprio usuário para suas **necessidades personalizadas**. Nessa opção, o usuário fornece exemplos de textos de um problema com o qual lida freqüentemente em suas buscas na Internet, e o sistema, através de marcadores estilísticos, gera um esquema de classificação novo para aquele usuário. Os exemplos devem ser de problemas binários (de duas classes), que estejam relacionados a tipos de texto, assim como as sete necessidades citadas anteriormente. Por exemplo, no caso de um advogado, distinguir entre textos técnicos sobre direito e textos voltados para o público comum, como a sentença dada para uma determinada ação e uma página informal sobre os direitos do consumidor, respectivamente. Essa abordagem não serve para problemas de classificação relacionados ao assunto, como, por exemplo, distinguir entre textos científicos que falam sobre problemas do coração da área de cardiologia e textos de outras áreas médicas que também falem sobre problemas do coração.

Para testar a opção personalizada foram inicialmente utilizados dois corpora, um criado em um mestrado do ICMC [4], e outro criado para o trabalho de doutorado para testes. O propósito do primeiro corpus é distinguir se uma página contém descrições de produtos à venda ou não e é composto por 1252 páginas (723 exemplos positivos e 529 negativos). O segundo corpus é composto por 200 páginas relacionadas ao domínio de direito; tem o propósito de distinguir entre páginas de direito para pessoas da área (advogados, juízes, etc.) e textos para pessoas em geral, sendo formado por 100 exemplos positivos e 100 negativos.

A avaliação da busca personalizada foi feita também com corpora criados por seis usuários, cinco portugueses e um brasileiro, dos quais dois têm formação em letras e quatro em computação. Foi solicitado a cada um por e-mail que descrevessem o problema que seria tratado por seus corpora e que criassem cada um, um corpus com 200 textos, sendo 100 exemplos positivos e 100 negativos. Foram criados sete corpora em resposta à solicitação por e-mail. Tanto os corpora[4] criados para o trabalho de doutorado, como o protótipo de ferramenta de busca na rede criado para utilização nos testes com os diversos esquemas de classificação[5] estão disponíveis no sítio da Linguateca. A descrição dos problemas tratados em cada um dos corpora personalizados é apresentada no Quadro abaixo.

Descrição informada pelos usuários para as sete necessidades personalizadas tratadas

Problema 1. Obter textos teóricos em html sobre filosofia da linguagem e sobre os principais pensadores e não textos (também em html) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico.

Problema 2. Obter textos teóricos em html sobre língua portuguesa e sobre os principais pensadores e não textos (também em html) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico.

Problema 3. Diferenciar textos que apresentem fatos sobre Fado de textos que emitam opiniões. No primeiro caso estão textos contendo informação histórica, biografias, notícias, etc. No segundo, entrevistas, críticas a discos e espetáculos, etc.

Problema 4. Encontrar textos que sejam uma descrição sobre determinado tema de História Geral. Entretanto, páginas com eventos, conferências, catálogos de livros não interessam, bem como informações sobre cursos de História, *links* para páginas de História ou ementas de disciplinas. Além disso, relatos de pessoas sobre seu gosto pela História também não são de interesse.

Problema 5. Distinguir glossários, receitas e técnicas sobre culinária japonesa de anúncios de livros, informação nutricional, críticas a restaurantes, páginas de restaurantes, informação sobre alimentação, cursos, festivais gastronômicos ou culturais sobre o mesmo tema.

Problema 6. Textos que interessam são história/fatos sobre surrealismo, como "Salvador Dali e o Surrealismo", "Manifesto do Surrealismo" e "Enciclopédia Universal Multimídia On-line". Blogs, exposições ou opiniões como "BdE - Blogue de Esquerda (II) 80 ANOS DE SURREALISMO", "Adelto Gonçalves,- comemorações", "A estranha sombra do surrealismo português, não interessam".

Problema 7. Documentos relevantes são: 1 - Documentos que explicam os princípios físicos que permitem que os aviões voem; 2 - Explicações técnicas de partes de componentes de aviões, tais como altímetros, tipos de motores ou rotores de helicópteros, etc.; 3 - História da aviação - biografia de pioneiros da aviação, os avanços aeronáuticos ao longo do tempo; 4 - História dos aviões - História de certos aviões importantes para a historia, as suas características, o motivo do seu desenvolvimento, o seu impacto na história da aviação. Documentos não-relevantes são: 1 - Notícias relacionadas com compras de aviões e empresas de aviação comercial; 2 - Notícias e descrições detalhadas de acidentes aéreos; 3 - Relatos de desvio de aviões e terrorismo aéreo; 4 - Opiniões sobre pilotagem, histórias e relatos de clubes de aviação, diversos documentos sobre psicologia do avião, deveres dos pilotos, etc.

[4] Os corpora estão disponíveis em http://www.linguateca.pt/Repositorio/YesUser/.
[5] O protótipo e seu código fonte encontram-se disponíveis em http://linguateca.pt/Repositorio/leva-e-traz/.

Cinco conjuntos de marcadores estilísticos foram utilizados para a criação dos diversos tipos de classificadores.[6] O primeiro foi criado com base nos trabalhos de Biber [1] e Karlgren [5]. Para sua criação foram consideradas intuições lingüísticas e foi dada preferência a marcadores que pudessem ser calculados sem a ajuda de qualquer tipo de analisador, como etiquetadores morfossintáticos e sintáticos. Esse conjunto contém 46 marcadores e é formado por estatísticas baseadas em palavras, como número de palavras longas; estatísticas baseadas no texto como um todo, como número de frases; e outras estatísticas, como número de advérbios de lugar.

O segundo conjunto de marcadores utilizado é composto por cinco funções para medir a riqueza de vocabulário propostas em [3].

O terceiro conjunto de marcadores é composto pelas 62 palavras mais freqüentes do corpus de necessidades: eliminando-se as *stopwords*, verbos auxiliares, advérbios, palavras relacionadas a domínios e agrupando algumas das palavras mais freqüentes como um único marcador. Esse conjunto é utilizado apenas nos experimentos com a classificação em sete necessidades de busca, pois possui marcadores dependentes dessa tarefa, como, por exemplo, número de ocorrência das palavras ”download” e “kb”.

Foi também utilizado um conjunto formado por 15 marcadores sintáticos selecionados com base em intuição lingüística, que foram calculados com ajuda do etiquetador sintático PALAVRAS [2]. Além de outro formado por 27 marcadores de aparência gráfica (layout). Exemplos de marcadores sintáticos são a porcentagem de sujeitos pronominais e o número de orações subordinadas e de marcadores de aparência gráfica são características de documentos html que indicam decisões como fonte utilizada, espaçamento e cor.

Referências

[1] Douglas Biber. Variation across Speech and Writing, Cambridge University Press, Cambridge, Inglaterra, 1988.

[2] Eckhard Bick. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

[3] Efstathios Stamatatos, George Kokkinakis e Nikos Fakotakis. Automatic text categorization in terms of genre and author. Em Computational Linguistics, Volume 26, 2000, 471 — 495.

[4] José Martins Junior e Edson dos Santos Moreira. Using Support Vector Machines to Recognize Products in E-commerce Pages. Em IASTED International Conference, fevereiro de 2004, 212—217.

[5] Jussi Karlgren. Stylistic Experiments for Information Retrieval. Tese de doutorado, Universidade de Estocolmo, 2000.

[6] Rachel Virgínia Xavier Aires. Uso de marcadores estilísticos para a busca na Web em português. Tese de doutorado, Universidade de São Paulo, ICMC, setembro de 2005.

[7] Sandra Maria Aluísio, Gisele Pinheiro, Marcelo Finger, Maria das Graças Volpe Nunes e Stella Esther Tagnin. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. Em Corpus Linguistics 2003, Lancaster, Inglaterra, 2003, v. 16, 14—21.

[6] A lista completa dos cinco conjuntos de marcadores utilizados é mostrada em [6].