UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS DE COMPUTAÇÃO E ESTATÍSTICA

O USO DE CARACTERÍSTICAS LINGÜÍSTICAS PARA A APRESENTAÇÃO DOS RESULTADOS DE BUSCA NA WEB DE ACORDO COM A INTENÇÃO DA BUSCA DO USUÁRIO

Rachel Virgínia Xavier Aires
ICMC-USP
raires@icmc.usp.br

Sandra Maria Aluisio
ICMC-USP
sandra@icmc.usp.br

Diana Santos
LINGUATECA / SINTEF ICT
Diana.Santos@sintef.no

Contexto

Inovações científicas, tecnológicas, culturais e sociais acontecem a todo tempo. Por ser uma mídia atualizada a cada segundo por diversas pessoas, a Internet tem o potencial de trabalhar como uma base de conhecimento, nos propiciando informações para atualizarmos nosso conhecimento e consequentemente nos adaptarmos às mudanças. Em sua forma textual, nos fornece os mais diversos tipos de texto, de vários gêneros e registros, como notícias, opiniões, instruções, receitas, reportagens, entrevistas, artigos, propaganda, livros, manuais, informações diversas e serviços. O tamanho da internet em português foi estimado em 5,090,230,228 palavras em 2002 [1].

Problema

Tanta infomação também traz problemas. Há 20 anos, contávamos com processos relativamente simples de filtragem feitos pelos editores de jornais, por exemplo, que selecionavam os artigos que seus leitores gostariam de ler. Hoje, esse tipo de barreira para informações inúteis não é mais tão eficiente. Desperdiçamos um grande número de horas procurando por informações e lendo informações que nunca utilizaremos, mesmo quando utilizamos sistemas de Recuperação de Informação (RI) como Google e Yahoo.

Propósito da Pesquisa

Este trabalho se propõe a definir uma abordagem para apresentação de resultados de sistemas para RI para português que resolva ou minimize consideravelmente o problema de usuários terem que lidar com um grande volume de documentos irrelevantes para ter acesso à informação desejada.

Metodologia de Pesquisa a ser utilizada

Após levantarmos as possibilidades de explorar características da língua portuguesa no processo de RI como um todo (veja o estudo relatado em [3]), decidimos tratar a apresentação de resultados, para apresentar de forma clara o enfoque dado por cada resultado para um dado tópico de consulta. Atualmente, estamos refazendo os experimentos relatados em [2] com mais dados de treinamento e explorando diferentes características da língua, para aumentarmos a precisão das regras de classificação de textos segundo o enfoque desejado. Nosso próximo passo será desenvolver o protótipo de uma ferramenta de busca na web para desktop para avaliarmos com usuários a eficiência de nossa abordagem.

Resultados Esperados

Pretendemos obter: (1) uma alternativa para apresentação de resultados que poupe o usuário do esforço de lidar com documentos que são relevantes segundo o tópico da consulta, mas não são para o usuário naquele instante; (2) o protótipo de uma ferramenta de busca para desktop para o português; (3) uma revisão da RI sob a perspectiva do PLN tendo como instância a língua portuguesa.

Além destes resultados diretos, o corpus criado para treinamento ficará disponível, podendo ser utilizado em outros trabalhos de classificadores, de RI, de Processamento de Língua Natural ou de Linguística Computacional, que precisem de corpus de textos extraídos da web, anotados com necessidades de usuários.

Agradecimentos

Os autores agradecem à FCCN (www.fccn.pt) que através da Fundação para a Ciência e Tecnologia (FCT) e co-financiada pelo POSI (POSI/PLP/43931/2001) financia este projeto desde setembro de 2001.

Referências

[1] AIRES, RVX; SANTOS, D. Measuring the Web in Portuguese. In: Euroweb 2002, December 2002, Oxford, UK, p. 198-199. Disponível em: http://www.nilc.icmc.usp.br/nilc/pessoas/rachel.htm. Acesso em: 13/10/2003.

[2] AIRES, RVX; MANFRIN, A; ALUÍSIO, SM; SANTOS, D. What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users'Needs. In: LREC 2004, may 2004, Lisbon, Portugal, p. 1943-1946.

[3] AIRES, RVX; ALUÍSIO, SM. Como incrementar a qualidade dos resultados das máquinas de busca: da análise de logs à interação em português. In: Revista Ciência da Informação, Vol 32, n. 1, p. 5-16, janeiro/abril 2003.

Artigos Publicados

[4] AIRES, RVX; ALUÍSIO, SM; QUARESMA, P; SANTOS, D.; SILVA, M. An initial proposal for cooperative evaluation on information retrieval in Portuguese. In: PROPOR (6th Workshop on Computational Processing of the Portuguese Language), june 2003, Faro, Portugal, p. 227-234. (c) Springer-Verlag.

[5] AIRES, RVX; ALUÍSIO, SM. Eu falo português. E daí? Poster in IHC 2002 (5th Symposium on Human Factors in Computer Systems), Fortaleza - CE, October, 2002.