Uso de marcadores estilísticos para a busca na Internet em português

Trabalho de doutorado desenvolvido entre setembro de 2001 e setembro de 2005.

Financiado por Fundação para Computação Científica Nacional (FCCN) através da Fundação para a Ciência e Tecnologia e co-financiada pelo POSI (POSI/PLP/43931/2001).

Objetivo

- 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

Solução explorada

- Classificar textos em gêneros, tipos de textos, necessidades de busca e necessidades personalizadas

Os classificadores...

- Foram criados utilizando:
  - Marcadores estilísticos
    - Utilizados em outros trabalhos para:
      - determinação de autoria
      - determinação da data de um determinado trabalho
      - escrita colaborativa
      - ensino de escrita através de sistemas tutores
      - geração de texto
      - tradução automática
      - classificação de textos segundo gêneros (escritos em alemão, inglês e grego)
  - Algoritmos de aprendizado de máquina
  - Corpora compilados com textos em português

Gêneros tratados

- Esquema de gêneros do Lácio-Ref:
  - Científico
  - Informativo
  - Jurídico
  - Literário
  - Instrucional

Taxa de acerto de 94,87%

Tipos textuais

- 29 tipos textuais do Lácio-Ref, por exemplo:
  - Reportagem
  - Artigo
  - Decreto
  - Monografia
  - Entrevista
  - Crônica
  - Relatório

  Taxa de acerto de 83,95%

Necessidades de busca

- Encontrar páginas que:
  - Definam alguma coisa ou ensinem como e/ou porque algo acontece
  - Ensinem como fazer algo ou como algo é feito
  - Forneçam uma apresentação (ou apanhado ou panorama) sobre um determinado assunto
  - Apresentem notícias
  - Forneçam informações sobre uma pessoa ou empresa ou instituição, ou organização
  - Página específica que o usuário quer visitar, mas não se lembra da URL
  - Forneçam algum serviço online

Taxa de acerto de 79,38%: 6 necessidades

Taxa de acerto de 91,19%: serviço x informação

Necessidades personalizadas

- Páginas de direito para pessoas da área (advogados, juizes, etc.) e textos para pessoas em geral (leigos): 83,2%
- Páginas que contém descrições de produtos a venda ou não: 89,91%
- 7 necessidades criadas por usuários: 87,5%-65,5%

Protótipo Leva-e-traz

Corpora classificados segundo necessidades e necessidades personalizadas:

- http://www.linguateca.pt/Repositorio/YesUser/

Leva-e-traz:

- http://www.linguateca.pt/Repositorio/leva-e-traz/

Tese:

- http://www.linguateca.pt/documentos/TeseDoutRachelAires.pdf