Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Aplicação de uma Técnica Tradicional de Expansão de Consulta ao Modelo TR+

Thyago Bohrer Borges
Vera Lúcia Strube de Lima

Janeiro de 2008

Agenda

- Introdução
- Expansão de Consultas
- Técnicas Tradicionais de EC
- Modelo TR+
- Proposta de Trabalho

Introdução

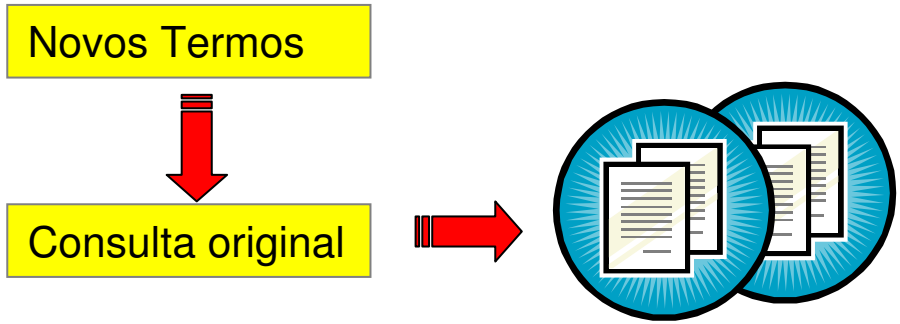
- **SRI's visam auxiliar o usuário a encontrar informações em um grande volume de dados**
- **SRI's (textual) - atende consultas dos usuários através de**
  - Indexação
  - Classificação
  - Busca de documentos

Introdução

- Formular uma consulta de forma eficiente não é um processo trivial
- Uma alternativa é o uso de técnicas de Expansão de Consultas
- Nos dias de hoje técnicas de PLN estão cada dia mais presente na EC

Expansão de Consultas - EC

- **Adicionar novos termos à consulta original**
  - Termos correlacionados
- **Porque?**
  - Aumentar a qualidade do processo de recuperação de informações (documentos)
    - Precisão e Abrangência



Técnicas Tradicionais de EC

- ***Relevance Feedback* (RF)**
  - Técnica de EC muito utilizada, implementação simples
  - Expõe ao usuário uma lista de informações relevantes a consulta original

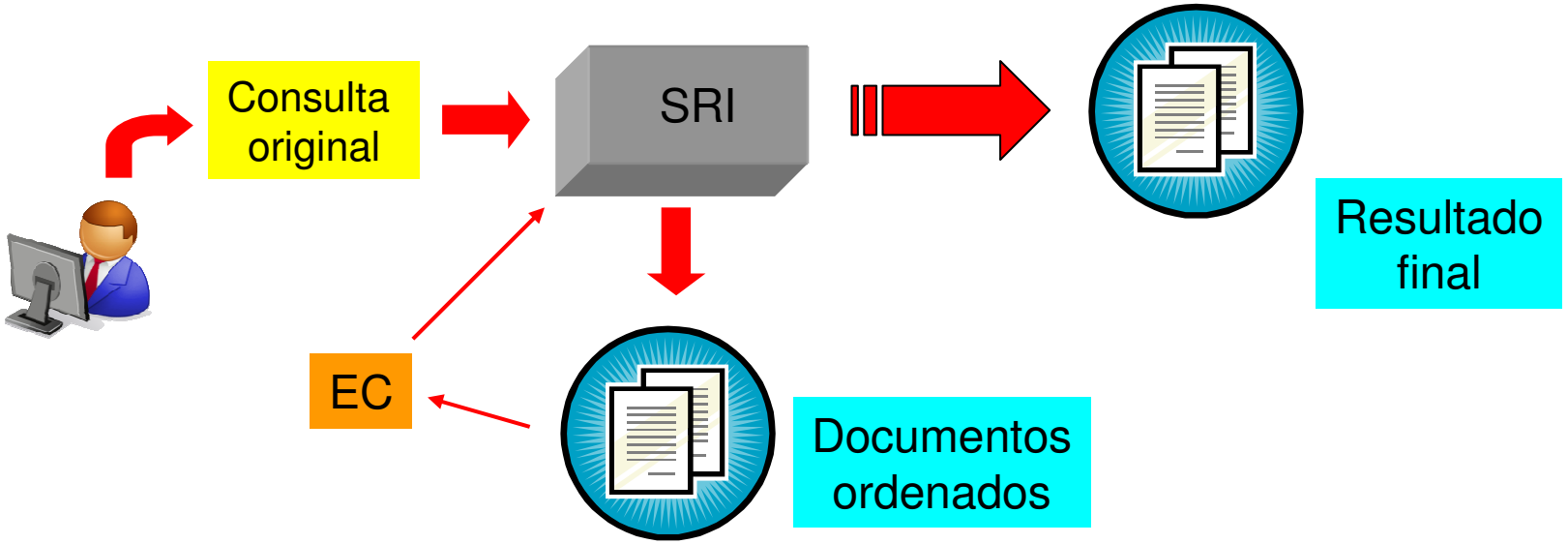
Lista de documentos

Lista de palavras

Snippets dos documentos

Técnicas Tradicionais de EC

- ***Pseudo Feedback***
  - Também conhecida como *Blind Feedback*
  - Variação da técnica *Relevance Feedback*
  - Os top *n* documentos recuperados são relevantes, e então reformulam a consulta



Técnica Tradicionais de EC

- **Análise Local Automática**
  - Após a consulta inicial, determina termos similares analisando os *n* primeiros documentos recuperados
  - Esta análise se baseia apenas em um conjunto "local" de documentos específico para uma consulta
  - Evita ambigüidade, uma vez que considera apenas documentos relevantes em um contexto

Técnicas Tradicionais de EC

- **EC baseada em *Thesaurus* de Similaridade**
  - Baseado no relacionamento *termo-a-termo*
  - Matriz de co-ocorrência
  - Cada termo é indexado ao documento onde ele aparece
  - Associação de agrupamentos pela co-ocorrência dos termos
  - Documentos são interpretados como elementos indexados

Técnicas Tradicionais de EC

- **EC baseada em *Thesaurus* Estatístico**
  - *Thesaurus* são produzidos de forma manual
    - Poucas línguas
    - Limitados no tipo de relações semânticas que representam
  - Termos com co-relação semântica podem ser descobertos utilizando análises estatísticas em uma coleção de documentos

Modelo TR+

- **Modelo TR+**
  - **SRI**
    - Indexação, classificação e recuperação dos documentos

❑ **Consulta**

- Termos Simples
  - Nominalização
    - Substantivo Abstrato
    - Substantivo Concreto
- Termos Compostos
  - RLB
    - Classificação
    - Restrição
    - Associação

Modelo TR+

- Nominalização

Palavra Original	Classe	Substantivo Abstrato	Substantivo Concreto
Correr	Verbo	Corrida	Corredor
Soldado	Particípio	Solda	Soldador
Sujo	Adjetivo	Sujeira	-

Modelo TR+

Relações Lexicais Binárias

Classificação	Restrição	Associação
=(subclasse ou instância,classe)	preposição(modificado,modificador)	evento(sujeito,objeto ou adjunto) evento.preposição(sujeito,objeto ou adjunto)
=(miau,gato)	de(time,futebol)	superacao(aluno,dificuldade)
=(gato,animal)	por(orientação,professor)	moradia.em(rainha,inglaterra)

Modelo TR+

- Documentos pré-indexados

Arquivo de índice no Modelo TR+

Termos ou RLB's	Id do documento	Peso	Id do documento	Peso
escola	A	0,65	B	1,49
sala	A	1,49	B	1,49
estudioso	A	1,22	B	1,22
fuga	A	1,74	B	1,64
trabalho	A	1,49		
=(aluno,fugitivo)	A	1,64	B	1,64
=(aluno,trabalhador)	A	1,98		
de(fuga,aluno)	A	2,00		
de(fuga,professor)	A	1,91	B	2,01
de(trabalho,professor)			B	2,00

Modelo TR+

- **Consulta Booleana**

Consulta original → Disjunção lógica de RLB's
ou
Conjunção lógica de Termos

Ex.

aluno estudioso → de(estudo,aluno) *ou* ≠de(estudo,aluno)
ou
((aluno) *e* (estudo ou estudante))

Classificação dos documentos recuperados →

Grupo superior : 1 ou + RLB's ou todos os termos

Grupo inferior : 1 ou n-1 termos

Trabalhos Recentes que utilizam EC

- **Uma Nova Proposta para Avaliação de EC: Má Combinação entre termos da consulta e dos documentos (Custis&Al-Kofahi,2007)**
  - EC combinação entre termos da consulta e os termos de uma coleção de um domínio específico
  - Avaliação usando fórmula de avaliação *OKAPI*
  - *Pseudo Feedback*
  - Mecanismos de Busca TCS
  - Utilização IDF para identificar os termos para EC

Trabalhos Recentes que utilizam EC

- **EC com termos selecionados usando Coesão lexical da análise de documentos (Vechtomova & Karamuftuoglu,2007)**
  - Ligações lexicais coesivas entre os termos da consulta e os termos do documento vizinhos aos termos da consulta no documento
  - Utiliza *snippets*
  - *Snippets* x Documentos inteiros
  - *Pseudo Feedback*

Trabalhos Recentes que utilizam EC

- **EC Personalizada para a Web (Chirita & Nejdl,2007)**
  - Termos para EC extraídos de um repositório pessoal do usuário
  - Compara os termos mais freqüentes com *WordNet*
  - Utiliza relações lexicais dos termos
  - EC com *Pseudo Feedback*

Proposta de Trabalho

- **Proposta**
  - Estudar técnicas de EC para ser agregada ao Modelo TR+
  - Utilização da técnica *Pseudo Feedback* ao Modelo TR+
  - Utilização dos Termos e RLB's na EC

Proposta de Trabalho

- **Questão de Pesquisa**
  - "A aplicação do método tradicional de expansão de consulta *Pseudo Feedback* ao Modelo TR+ pode representar um ganho de precisão e abrangência na recuperação dos documentos?"

Proposta de Trabalho

- **Partindo-se desta questão de pesquisa, tem-se a seguinte hipótese:**
  - A utilização da técnica de EC *Pseudo Feedback*, auxilia na recuperação de informações, aumentando a precisão e a abrangência das informações recuperadas pelo Modelo TR+

Proposta de Trabalho

- **Validação da proposta**
  - Termos Nominalizados
    - Substantivos Concretos
    - Substantivos Abstratos
  - Utilização de RLB's na EC
  - Tipos de RLB's na EC
    - Restrição
    - Associação
    - Classificação
  - Precisão, Abrangência....

Muito obrigado!

thyago.borges@gmail.com

vera@inf.puc.br

http://www.inf.pucrs.br/~linatural