

Operações sobre Texto

Nem todas as palavras num texto não igualmente importantes para representá-lo semanticamente. Geralmente substantivos (ou grupos de substantivos) são mais representativos do conteúdo de um documento.

Usar o conjunto de todas as palavras numa coleção para indexar seus documentos gera muito **ruído** para a tarefa de recuperação. Para reduzir o ruído deve-se diminuir o conjunto de palavras usadas para representar o documento.

O pré-processamento dos documentos numa coleção simplesmente controla (diminuindo) o tamanho do vocabulário, isso leva a um melhor desempenho nos SRI.

Problema: Usuários podem usar palavras que não aparecem nos documentos devido a este controle, por exemplo, uma consulta com os termos “A casa dos espíritos”, documentos que até poderiam ter frases idênticas a esta, pelo fato dos termos “A” e “dos” não aparecerem no vocabulário, poderiam não ser retornados.

É importante que fique claro para o usuário do SRI que tipos de palavras ele deve usar em suas consultas.

Por este problema, máquinas de busca na *web* costumam ignorar o pré-processamento e fazer a indexação *full text*, pois apesar do ruído, é mais eficiente para usuários leigos (da internet).

Outras formas de melhorar o desempenho dos SRI, além do pré-processamento:

- A construção de Tesauros para representar relacionamentos conceituais entre as palavras;
- *Clustering* (agrupamento) de documentos relacionados;
- Compressão de textos;
- Criptografia; dentre outros.

Estas técnicas melhoram o desempenho dos SRI no que diz respeito à precisão e não à velocidade na resposta (muito pelo contrário). As máquinas de busca na *web* costumam não implementar estas técnicas devido a necessidade de respostas rápidas.

Pré-processamento dos Documentos

O pré-processamento pode ser dividido basicamente em cinco operações sobre textos:

- (1) Análise Léxica
- (2) Eliminação de *stopwords*
- (3) *Stemming* das palavras restantes
- (4) Seleção de termos de indexação
- (5) Construção de estruturas de categorização de termos

Análise Léxica

A análise léxica é a etapa responsável por identificar palavras no texto. Tem como objetivo tratar espaçamento, hifenização, pontuação, caracteres especiais, etc.

Eliminação de *stopwords*

Esta etapa tem como objetivo filtrar palavras com valores discriminatórios baixos para a tarefa de recuperação. Também pode ser considerada uma técnica de compressão de textos. Pode diminuir o tamanho do texto em até 40%.

É elaborada uma lista de *stopwords*, formada por palavras tais como artigos, conjunções, pronomes, e que podem chegar a conter verbos, adjetivos, etc. Um exemplo desta lista contém 425 palavras em inglês.

Observação importante: o uso de *stopwords* pode reduzir o *recall*.

Stemming

O *stemming* (corte) de palavras restantes tem como objetivo remover prefixos e sufixos, permitindo a recuperação de variações sintáticas das palavras. Um *stem* é a parte que resta de uma palavra quando são retirados seus afixos.

Existem conflitos na área sobre o fato do *stemming* trazer melhorias de desempenho. Estes conflitos são observados em experimentos.

Seleção de Termos de Indexação

Esta etapa tem como objetivo determinar que palavras/*stems* (ou grupos de palavras) serão utilizadas como elementos de indexação (substantivos são mais representativos que adjetivos, verbos, advérbios, etc.).

Podem ser usadas todas as palavras ou escolher algumas mais significativas. Pode ser feita de forma manual ou automática.

Uma das técnicas automáticas, usada no INQUERY, separa substantivos, pronomes, verbos, adjetivos e advérbios em uma sentença.

Construção de Estruturas de Categorização de Termos

Estas construções, tais como Tesauro, tem como objetivo extrair estruturas diretamente representadas no texto, para permitir, por exemplo, a expansão de consultas.

Tesauro é uma lista pré-compilada de palavras importantes num determinado domínio de conhecimento. E, para cada palavra da lista, é criado um conjunto de outras palavras relacionadas.