E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

# TÉCNICAS APLICADAS EM MÁQUINAS DE RECUPERAÇÃO TEXTUAL

## *TECHNIQUES APPLIED IN TEXTUAL SEARCH ENGINES*

**Raphael Winckler de Bettio**
Summa Technologies do Brasil, E-mail: raphaelwb@gmail.com

**Alejandro Martins Rodriguez**
Sociedade Educacional de Santa Catarina (SOCIESC), E-mail: aljmartins@gmail.com

**Resumo:** Desde a criação dos computadores o montante de informação arquivada vem crescendo exponencialmente, em consequência foi necessária a criação de ferramentas computacionais que permitissem a recuperação de informação útil em meio às bases de armazenamento. Com esse objetivo em 1950 surgiu um novo campo denominado recuperação textual. O principal objetivo deste artigo é apresentar um resumo das técnicas que fazem parte deste campo de pesquisa.

**Palavras-chave**: Extração de termos, Expansão de query, Busca textual, Ontologias, Semântica.

**Abstract:** Since the introduction of computers the amount of stored information is growing exponentially, and, as a result, it was necessary to create computational tools that would allow the recovery of useful information stored inside databases. With this objective in 1950 was created a new field called textual recovery. The main objective of this paper is to present a summary of techniques that are part of this research's field.

**Key-words:** Term extraction, Query expansion, Textual search, Ontology, semantics.

## 1 INTRODUÇÃO

> "Durante toda a história da humanidade, a cada novo paradigma que aparece voltamos a zero em termos dos padrões e regras até então utilizados. Neste processo evolutivo, passamos da Sociedade Agrícola para a Sociedade Industrial e da Sociedade da Informação e para a Era do Conhecimento, a qual está baseada no conhecimento e em valores intangíveis que este conhecimento poderá trazer de retorno às organizações" (RODRIGUEZ y RODRIGUEZ, 2001, p. 05).

Segundo Drucker (1994 apud Ponchirolli, 2000) um novo contexto sócio-econômico denominado Sociedade do Conhecimento está surgindo. Essa nova sociedade está caracterizada, principalmente, por um período de rápidas mudanças tecnológicas, econômicas e sociais (DRUCKER *apud* PONCHIROLLI, *2000)*. Segundo Crawford (1994), os próximos anos nos reservam um período em que empresas seculares desaparecerão

**E-Tech:** Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

em um ano, um período onde países em que ninguém acreditava começarão a emergir como novas forças mundiais. Essas mudanças vêm surgindo em função de uma profunda transformação na economia mundial. Enquanto países de terceiro mundo passam pelo processo de industrialização, as economias desenvolvidas são rapidamente transformadas em economias pós-industriais baseadas em conhecimento.

Como parte integrante desse processo e com a finalidade de amparar a nova sociedade em seu desenvolvimento surgiu a Engenharia do Conhecimento. A Engenharia do Conhecimento tem como principal objetivo pesquisar acerca do conhecimento em todos os seus aspectos.

A forma de conhecimento que será abordado neste artigo é o manifestado através da escrita.

Conforme Sighal (2001) as pessoas sabem sobre a importância do armazenamento e busca de informação. Com o advento dos computadores, tornou-se possível o armazenamento de grandes quantidades de informação em bases de dados e, em consequência, catalogar a informação dessas bases tornou-se imprescindível.

Devido a essa necessidade, durante a década de 50 um novo campo do conhecimento surgiu, o campo da Recuperação da Informação. A partir do surgimento deste novo campo surgiram modelos computacionais capazes de tornar essa atividade possível. Este artigo apresenta as técnicas que são consideradas básicas para qualquer sistema de busca de informação atualmente desenvolvido:

a) *inverse Document Frequency:* técnica largamente utilizada, criada com o objetivo de verificar a relevância entre termos de bases de dados textuais;

b) *stopWords:* técnica que visa remover termos pouco significativos para melhorar o poder de processamento dos algoritmos;

c) *stemming:* utiliza como base conhecimentos da área linguística e tem como principal finalidade tornar possível aos algoritmos reconhecer a semelhança entre palavras;

d) *term Extration:* possibilita a seleção de termos que melhor representam um determinado documento em uma base de dados;

e) *query Expansion:* tem por objetivo melhorar as *Querys* (conjunto de palavras-chaves) informadas pelos usuários no momento da busca.

E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

## 2 INVERSE DOCUMENT FREQUENCY

Karen Sparck Jones publicou em 1972 no *Journal of Documentation* um artigo intitulado *"A statistical interpretation of term specificity and its application in retrieval"*. A medida proposta no artigo veio a ser conhecida como *Inverse Document Frequency* ou IDF e é baseada na contagem de um determinado termo em uma determinada coleção de documentos. (JONES, 1972).

A idéia baseava-se no pressuposto que um termo que aparece em muitos documentos não é um termo que representa bem um determinado documento, e a medida proposta era uma implementação heurística desse conhecimento.

A medida proposta por Jones (1972), atribuindo um peso ao termo é essencialmente, conforme expressão:

$$IDF(t_i) = \log\left(\frac{N}{n_i}\right)$$

Expressão 1: Fórmula do IDF
Fonte: Jones (1972)

A fórmula apresentada assume que N é o número de documentos em uma coleção, e o termo $t_i$ ocorre em $n_i$ documentos desta base. Salienta-se que termo não pode ser considerado uma palavra, uma frase ou uma *word stemming*.

O objetivo da utilização do logaritmo é garantir que, conforme a frequência de um termo aumente, sua importância em relação a frequências menores seja atenuada.

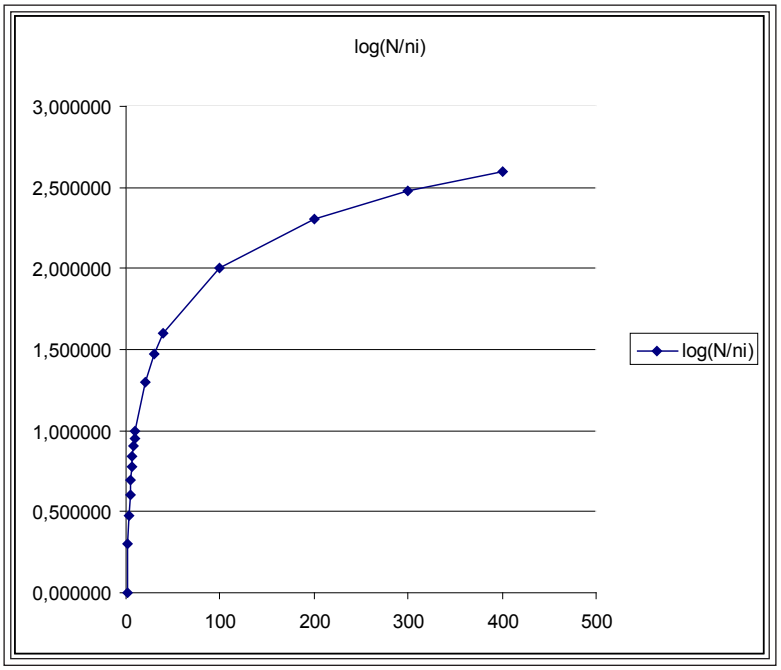


Figura 1: Representação Gráfica da Curva do IDF

De acordo com a tabela a seguir (Quadro 1), um termo com relevância 200 e um termo com relevância 300 tem praticamente o mesmo IDF e pressupõe-se que, depois de um determinado número de ocorrências, o termo perca sua relevância, e a curva que representa essa perda é a logarítmica (Figura 1)

| N/ni | log(N/ni) |
|---|---|
| 1 | 0,0000000 |
| 2 | 0,3010300 |
| 3 | 0,4771213 |
| 4 | 0,6020600 |
| 5 | 0,6989700 |
| 6 | 0,7781513 |
| 7 | 0,8450980 |
| 8 | 0,9030900 |
| 9 | 0,9542425 |
| 10 | 1,0000000 |
| 20 | 1,3010300 |
| 30 | 1,4771213 |
| 40 | 1,6020600 |
| 100 | 2,0000000 |
| 200 | 2,3010300 |
| 300 | 2,4771213 |
| 400 | 2,6020600 |

Quadro1: Representação Numérica da Curva do IDF

E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

Para uma melhor explicação sobre o funcionamento da medida de Jones (1972), a seguir estão relacionados trechos retirados da Enciclopédia Digital Wikipédia.

Trecho 1:

> A **Copa do Mundo**, ou Campeonato do Mundo de Futebol é um torneio de **futebol** masculino, realizado a cada quatro anos pela FIFA. Começou em 1930, com a vitória da seleção do Uruguai. No primeiro mundial, não havia torneio eliminatório, e os países foram convidados para o torneio. A Itália sagrou-se bicampeã em 1934 e 1938. Nos anos de 1942 e 1946, a Copa não ocorreu devido à Segunda Guerra Mundial. Em 1950, o mundial foi realizado no **Brasil**, que chegou como favorito à sua primeira final de Copa, mas a Celeste Olímpica uruguaia estragou a festa de 200 mil pessoas presentes no Maracanã, então o maior estádio do mundo, vencendo o jogo por 2x1, quando o empate teria sido suficiente ao **Brasil** para conquistar o título. O episódio ficou conhecido como Maracanazo. (WIKIPEDIA, 2005).

Trecho 2:

> O **Brasil** possui a seleção com mais títulos mundiais, o único país pentacampeão e o único a ter vencido o torneio fora do seu continente. É também o único país a participar de todas as Copas. Seguem-se as seleções tricampeãs da Alemanha e da Itália, as bicampeãs da Argentina e do Uruguai e, por fim, as seleções da Inglaterra e da França, com um único título. (WIKIPEDIA, 2005).

Trecho 3:

> A **Copa do Mundo** é o segundo maior evento esportivo do mundo, ficando atrás apenas dos Jogos Olímpicos de Verão. É realizada a cada quatro anos, tendo sido sediada pela última vez, em 2002, no Japão e na Coréia do Sul, com o **Brasil** como campeão. A próxima, em 2006, será na Alemanha. (WIKIPEDIA, 2005).

A Quadro 2 apresenta uma contagem do número de ocorrência de determinados termos nos trechos selecionados. Para este exemplo uma palavra ou um conjunto de palavras é considerado um termo. Com o objetivo de melhorar a precisão do modelo, pode-se utilizar a técnica de *StopWords*, assim, o termo "o" seria removido dos trechos textuais. Porém, aqui esta técnica não será utilizada, pois desta maneira a diferença entre os IDF serão maiores, simplificando, assim, seu entendimento. Na segunda coluna é apresentado o cálculo do IDF simplificado (N/ni), sem a utilização do *log*, e na terceira coluna o *log* é acrescentado.

| Termo | Quantidade | N/ni | log(N/ni) |
|---|---|---|---|
| futebol | 1 | 3,00 | 0,47712126 |
| Copa do Mundo | 2 | 1,50 | 0,17609126 |
| Brasil | 4 | 0,75 | -0,12493874 |
| O | 12 | 0,25 | -0,60206000 |

Quadro 2: Termos e Respectivos Pesos IDF

Seguindo as idéias de Jones (1972), é possível afirmar que o termo futebol representa melhor o trecho 1 do que o termo o, pois aparece apenas neste documento. Os valores calculados a partir da fórmula base representam esta afirmação. Já o termo Copa do Mundo é mais útil na classificação dos documentos que o termo o, pois aparece em apenas dois dos trechos, trecho 1 e trecho 3. Já o termo Brasil tem uma representatividade menor que os dois outros termos citados já que aparece em todos os trechos, mesmo assim possui uma representatividade melhor que o termo o.

Apesar de simples, as idéias de Jones (1972) são úteis e largamente utilizadas em conjunto com outras técnicas no que se refere a busca de documentos.

## 3 STOPWORDS

Apenas uma pequena parte das palavras contidas em um texto reflete a informação contida no mesmo. Analisando a língua inglesa é possível afirmar que palavras como *it*, *and* e *to* podem ser encontradas em praticamente qualquer sentença. Portanto, são termos que são extremamente pobres no que se refere a busca por documentos. Entretanto, representam a maioria dos termos dos documentos, estas palavras são conhecidas como *StopWords* (RACHEL, 2006).

A remoção de *StopWords* é uma tarefa existente em praticamente todos os sistemas de recuperação e informação textual. Uma lista de palavras consideradas *StopWords* pode ser construída analisando os textos que serão utilizados como base para a busca ou fazendo uma análise do idioma utilizado. A título de exemplo, relaciona-se uma lista de *StopWords* no quadro 3 utilizada pelo algoritmo de Porter (PORTER, 1980).

| | | | | | | |
|---|---|---|---|---|---|---|
| de | é | As | nos | eu | depois | eles |
| a | com | Dos | já | também | sem | estão |
| o | não | como | está | só | mesmo | você |
| que | uma | mas | seu | pelo | aos | tinha |
| e | os | Foi | sua | pela | ter | foram |
| do | no | Ao | ou | até | seus | essa |
| da | se | Ele | ser | isso | quem | num |
| em | na | das | quando | ela | nas | nem |
| um | por | tem | muito | entre | me | suas |
| para | mais | à | há | era | esse | meu |
| qual | essas | tu | minhas | nossa | estes | isto |
| será | esses | te | teu | nossos | estas | aquilo |
| nós | pelas | vocês | tua | nossas | aquele | havia |
| tenho | este | vos | teus | dela | aquela | seja |
| lhe | fosse | lhes | tuas | delas | aqueles | pelos |
| deles | dele | meus | nosso | esta | aquelas | elas |
| numa | têm | minha | às | | | |

Quadro 3: Stopwords Utilizados Pelo Algoritmo Porter

E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

## 4 STEMMING

Conforme Dennis (2000) no contexto da Recuperação da Informação, *Stemming* refere-se ao processo de remoção dos prefixos e sufixos das palavras. *Stemming* é usado para reconhecer os padrões de formação das palavras com a finalidade de recuperar a informação. Como um simples exemplo, considere a busca por um documento intitulado "Como Escrever". Se o usuário digitar "Escrevendo" o sistema não conseguirá encontrar nenhum documento, no entanto, se a entrada de dados for *stemmizada*, Escrever tornar-se-á "Escrev" – denominado *stem* – e o documento "Como Escrever" será apresentado ao usuário.

Diversos algoritmos foram desenvolvidos para este fim, cita-se Hooper e Paice (2005), Lovins (1968), Krovetz (1993). Porém, o algoritmo mais comumente utilizado é o algoritmo de Porter, escrito em 1980 e publicado no artigo "*An algorithm for suffix stripping*" (PORTER, 1980).

Para Porter (1980), uma versão modificada do algoritmo *Porter* denominado *Porter2* ou *SnowBall*, trata-se de uma versão melhorada do algoritmo original e deve ser utilizada para se obter melhores resultados.

O algoritmo *SnowBall* foi modificado para funcionar em diversos idiomas inclusive o Português, entretanto, algoritmos desenvolvidos especificadamente para um idioma tendem a apresentar melhores resultados, como é o caso do algoritmo publicado por Orengo (2001) e que será utilizado nesta pesquisa por apresentar melhores resultados que o algoritmo Porter.

Segundo Orengo (2001), os resultados apresentados pelo algoritmo desenvolvido especificadamente para a língua portuguesa apresentou um menor valor referente ao erro de *understemming* (redução de palavras com significados iguais para *stemmings* diferentes) e *overstemming* (redução de palavras com significados diferentes para o mesmo *stemming*) concluindo-se que o algoritmo é mais eficiente que o algoritmo Porter.

O algoritmo é composto por oito etapas, que estão apresentadas na Figura 2:

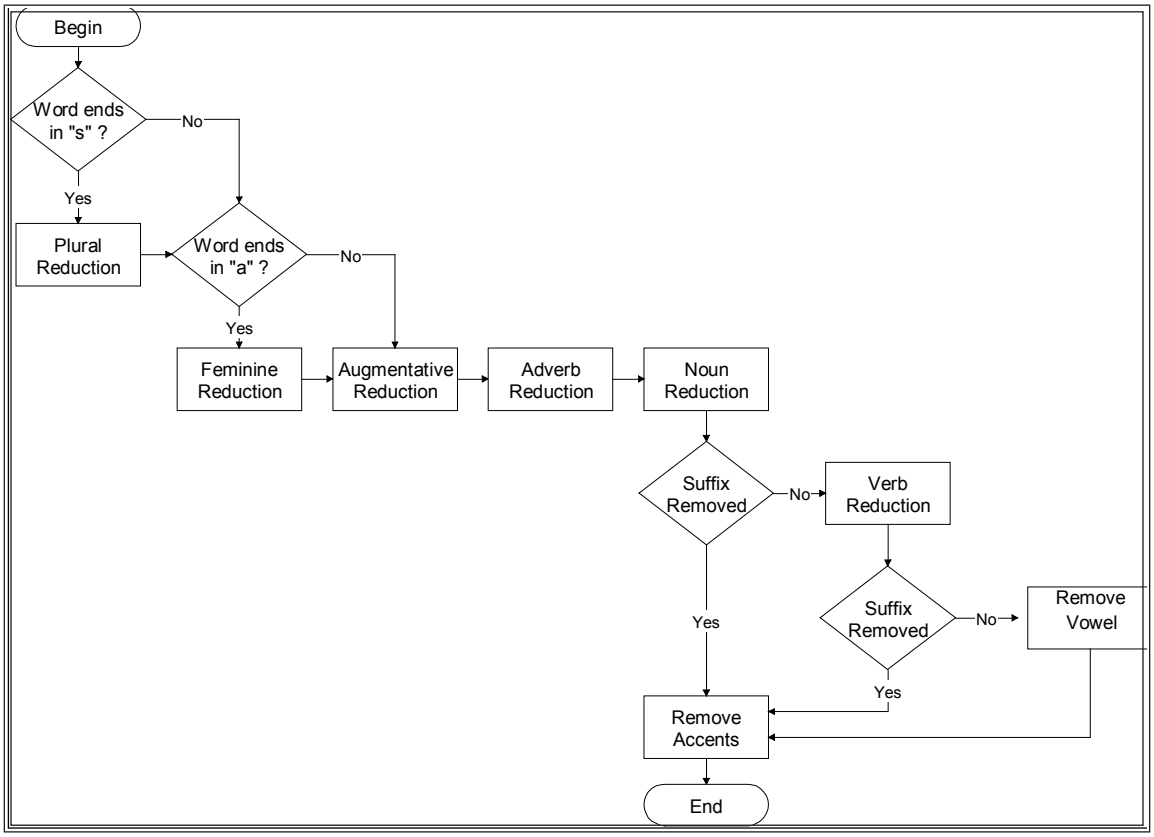


Figura 2: Fases do Algoritmo Stemming para Língua Portuguesa
Fonte: Orengo; Huyck (2001).

Cada passo do algoritmo é composto por um grupo de regras, sendo que cada passo é executado conforme o fluxo (Figura 2) e apenas uma regra pode ser executada por vez. O algoritmo possui 199 regras. Cada regra é composta de 4 fatores, a saber:

a) o sufixo a ser removido;

b) o tamanho mínimo do stem: este fator impede que determinados sufixos sejam removidos;

c) um sufixo que deve substituir o sufixo atual da palavra;

d) uma lista de excessões, palavras que apesar de cumprirem as regras não devem ser afetadas.

Para melhor compreensão do mecanismo das regras, segue o exemplo (Figura 3):

| “inho”, 3, ””, {“caminho”, “carinho”, “cominho”, “golfinho”, “padrinho”, “sobrinho”, “vizinho”} |
|---|

Figura 3: Exemplo de Regra do Algoritmo de Stemming da Língua Portuguesa

**E-Tech:** Tecnologias para Competitividade Industrial, Florianópolis, v. 2, n. 1, p. 33-48, 1º. Sem., 2009.

Onde “inho” é o sufixo a ser analisado, 3 é o tamanho mínimo do stem, o que previne que palavras como “linho” sejam stemmizadas, e uma lista de excessões onde esta regra não pode ser aplicada (carinho, caminho, etc.).

As 8 (oito) etapas do algoritmo estão descritos a seguir:

a) redução do plural: com raras excessões, a remoção do plural na língua portuguesa consiste na remoção da letra “s”;

b) redução do feminino: todos os Substantivos e Adjetivos na língua portuguesa possuem uma versão masculina. Esta etapa consiste em transformar a forma femina na forma correspondente masculina;

c) redução dos ddvérbios: esta etapa consiste em analisar palavras finalizadas em “mente”, como nem todas as palavras terminadas neste sufixo representam advérbios uma lista de excessões existe;

d) aumentativo/diminutivo: a língua portuguesa apresenta uma variação muito grande de sufixos utilizados nestas formas, entrentanto, apenas os mais comuns são utilizados para evitar o *overstemming*;

e) redução dos substantivos: esta etapa testa as palavras, procurando por 61 sufixos utilizados em substativos, se este sufixo é removido as etapas 6 e 7 são ignoradas;

f) redução dos verbos: a língua portuguesa é muito rica em termos de formais verbais, enquanto a língua inglesa possui apenas quatro variações, a língua portuguesa contém cinquenta diferentes formas;

g) remoção de vogais: esta etapa consiste em remover as letras “a” e/ou “o” no final das palavras que não tenham sido *stemmizadas* pelos passos 5 e 6;

h) remoção de acentos: a remoção de acentos é importante, já que existem palavras em que as mesmas regras se aplicam a versões acentuadas e não acentuadas (por exemplo, psicólogo e psicologia).

Ainda segundo Orengo (2001), os maiores dificultadores no que se refere a stemming da língua portuguesa são a quantidade de excessões nas regras, quantidade de palavras com mais de um significado, a quantidade de verbos irregulares, quantidade de palavras onde a raiz da mesma é alterada e a dificuldade em reconhecer nomes próprios.

A Quadro 4 apresenta alguns exemplos de palavras stemmizadas.

| Termo | Termo após Alg |
|---|---|
| bobalhões | Bobalhõ |
| bocadinho | Bocadinh |
| quintuplicou | Quintuplic |
| quimioterápicos | Quimioteráp |
| quilométricas | Quilométr |
| bocaiúva | Bocaiúv |
| quiosque | Quiosqu |

Quadro 4: Palavras Stemmizadas

## 5 TERM EXTRATION

A Extração de Termos (Term Extraction) é um importante problema a ser estudado, no que diz respeito ao processamento de linguagem natural. Seu objetivo é a extração de coleções de palavras que representem o significado de um texto. Muitos linguistas têm argumentado que a base semântica de um texto pode ser representada por estes termos. Diversos tipos de aplicações utilizam as técnicas de Term Extraction, entre elas pode-se citar: máquinas de tradução, ferramentas de indexação de documentos, construtoras de bases de conhecimento e sistemas de Recuperação da Informação (PANTEL e LIN, 2006).

Segundo Milios, Zincir-Heywood e Zhang (2005), o estado da arte dessas técnicas é figurado, atualmente, pelos algoritmos C-value/NC-value desenvolvidos por Frantziy (FRANTZIY, ANANIADOUY, MIMAZ, 2000), publicado no International Journal on Digital Libraries Manuscript .

Muitas técnicas para extração de termos foram criadas, entretanto, as técnicas criadas por Church e Dagan (1995), Justeson e Katz (1995),Enguehard e Pantera (1995) utilizavam apenas informação estatística. A técnica C-value/NC-value apresenta uma nova visão sobre o tema, combinando técnicas estatísticas com técnicas lingüísticas (FRANTZIY, ANANIADOUY, MIMAZ, 2000).

C-value é uma técnica de extração estatística mais eficiente que as já existentes e a técnica NC-value foi desenvolvida para incorporar informação contextual às informações já encontradas através da C-value. Seu algoritmo utiliza-se de informações estatísticas e linguísticas.

Basicamente, esse método tem como entrada um texto e como saída uma lista de termos candidatos ordenados pelo valor C-value, também denominado *termhood*. Essa lista deve ser analisada por especialista de domínio (domínio do texto utilizado). Não existe a necessidade de analisar todos os termos, no entanto, a eficiência do algoritmo está diretamente ligada à quantidade de termos analisados pelo especialista.

A parte linguística do algoritmo está baseada em uma lista de *StopWords* e um filtro linguístico que analisa o tipo (verbo, pronome, etc.) dos termos a serem extraídos. Salienta-se que um termo pode ser composto de uma ou mais palavras. A parte linguística consiste em três etapas:

a) marcar o texto com os tipos de termos;

b) aplicar um filtro linguístico que remove os tipos de termos indesejáveis;

c) excluir os termos pertencentes a uma *StopList* também conhecida por *StopWords*.

A parte estatística consiste na análise de quatro informações relativas a cada:

a) a frequência que o termo aparece no documento;

b) a frequência que o termo aparece em conjunto com outro termo do documento;

c) o número de termos a serem selecionados;

d) a quantidade de caracteres do termo.

De acordo com Frantziy, Ananiadouy e Mimaz (2000), o uso de informações estatísticas que vão além da simples frequência do termo da extração em conjunto com o uso de informações linguísticas aumenta significativamente a eficiência.

## 6 QUERY EXPANSION

Conforme BillerBeck e Zobel (2006) as máquinas de busca são o principal mecanismo usado para procurar documentos na Internet. Essas ferramentas utilizam mecanismos de Recuperação da Informação para comparar *queries*, expressas em uma série de palavras, com os documentos e julgar quais deles são melhores para responder a pergunta dos usuários.

Quando as *queries* são bem formuladas, consistindo em palavras-chave de um tópico específico, as quais juntas demonstram a informação necessária com um nível baixo de ambiguidade, as máquinas de busca conseguem obter documentos que refletem as palavras. Todavia, a maioria das *queries* não são bem formuladas, elas são ambíguas, não precisas o suficiente, ou usam os termos específicos para um determinado contexto. A maioria (60%) das *queries* imputadas nas máquinas de busca é formada por duas a três palavras. Esse tipo de entrada acaba trazendo como resultado uma quantidade grande de documentos, o que dificulta a análise.

Hoje, uma variedade de técnicas para aumentar a eficiência desses mecanismos é usada, uma delas é a *Query Expansion*. É possível afirmar que existem dois grupos básicos de técnicas utilizadas para expansão de *queries* (GROOTJEN, 2006), são elas:

a) user feedBack relevance: esta é provavelmente a técnica mais comum de reformulação de *queries*. Esta técnica requisita que o usuário atribua relevância a um conjunto de documentos trazidos através de uma busca inicial. Experimentos recentes têm demonstrado uma melhora significativa no resultado das buscas, trabalhando em bases com poucos documentos. O modelo matemático é simples de implementar, a única dificuldade com a técnica é persuadir o usuário a atribuir relevância a documentos, o que é um trabalho tedioso;

b) global query expansion: outra forma de expandir *queries* é adicionando palavras (sinônimos ou palavras relacionadas) à *Query* original. Para fazer isso, utiliza-se um *thessaurus* ou outro tipo de fonte de dados. *Thessaurus* são freqüentemente utilizados em sistemas de Recuperação da Informação como um mecanismo para reconhecer expressões sinônimas e entidades lingüísticas que são semanticamente similares, mas superficialmente distintas. Diferente da técnica de relevância através de *feedback*, não é necessário analisar a base dos textos.

Segundo Mandala, Tokunaga e Tanaka (1999), técnicas de expansão de *queries* utilizando-se *thessaurus* são alvos de pesquisas por quatro décadas e uma quantidade enorme de métodos foi desenvolvida. Os vários métodos podem ser enquadrados em três grupos básicos: *Hand-crafed thessaurus based*, *Co-occurrence-based automatically constructed thessaurus based* e *Head-modi er-based automatically constructed thessaurus based*.

A *Query Expansion* baseada em *Hand-Crafed Thessaurus* somente tem sucesso se o domínio do *thessaurus* é o mesmo domínio das bases textuais. De acordo com os experimentos da *Text Retrieval Conference* – TREC – o uso de *thessaurus* genéricos não tem tido muito êxito. Já o modelos que utilizam *thessaurus* construídos automaticamente (são construídos baseados em uma coleção de textos, sem intervenção humana) têm obtido pequenas taxas de eficiência, em torno de 20%.

Em 1992, foi criada a TREC – *Text Retrieval Conference – co-patrocinada pelo National Institute of Standards and Technology* – NIST – e pelo *U.S. Department of Defense*. Ela tem por finalidade realizar pesquisas no que diz respeito à recuperação de informações em bases textuais, e atualmente é considerado o mais importante congresso na área. Ao realizar uma análise de seus *proceedings* é possível assegurar que pesquisas na área de expansão de *queries* estão longe de se esgotar e diversas técnicas que utilizam cruzamento de métodos já criados foram especificadas com o objetivo de promover melhoria na eficiência das buscas.

Em seguida, citam-se algumas pesquisas desenvolvidas, que utilizam *Query Expansion*:

a) UMass Robust 2005. Using Mixtures of Relevance Models for Query Expansion: utiliza como base as técnicas de aproximação de termos e também técnicas de pseudo relevância através de *feedback;*

b) symbol-Based Query Expansion Experiments: estuda a eficiência de algoritmos de expansão de *Querys* baseados no *feedback* dos usuários;

c) the Effects of Primary Keys, Bigram Phrases and Query Expansion on Retrieval Performance: procura expandir as *Querys* através da análise estatística dos termos contidos na base textual utilizada;

d) concept-Based Query Expansion and Bayes Classification: utiliza uma máquina de indexação de conceitos denominada Collexis para expandir as *queries*.

Um exemplo de query expansion baseado em thessaurus está abaixo citado:

a) thessaurus: Automóvel = Carro, Roubo = Furto, CD-Player = Som;

b) query formulada pelo usuário: O Som do Carro foi roubado;

c) query reconstruida através de Query Expansion: O Som CD-Player do Carro Automóvel foi Roubado Furto.

## 7 CONCLUSÕES

O artigo apresentou uma série de técnicas que podem ser aplicadas em ferramentas que visem a busca de dados em bases textuais. As técnicas aqui apresentadas podem ser utilizadas em conjunto ou isoladas dependendo da necessidade de cada sistema de busca. É importante salientar que as técnicas apresentadas possuem variantes que foram construídas para melhorar performance ou com o objetivo de serem aplicadas para resolver problemas específicos, neste artigo procurou-se apresentar as mais comuns destas variantes.

Assim sendo conclui-se que os objetivos do artigo no que tange a apresentação de técnicas básicas de busca textual foi abordado, entretanto, deixa-se em aberto uma lacuna onde poderia ser tratado sobre o uso em conjunto destas técnicas.

## REFERÊNCIAS

BILLERBECK, B.; ZOBEL, J. **Questioning query expansion**: an examination of behavior and parameters. School of Computer Science and Information Technology. Australia: RMTI University, 2006.

CHURCH, K.; DAGAN, I. **Termight**: identifying and translating technical terminology. USA: AT&T Bell Laboratories, 1995.

CRAWFORD, R. **Na era do capital humano**. São Paulo: Atlas, 1994.

DENNIS, S. **Stemming algorithms for information retrieval and question/answer systems**. Colorado: Institute of Cognitive Science University of, 2000.

ENGUEHARD, C.; PANTERA, L. Automatic natural acquisition of a terminology. **Journal of Quantitative Linguistics**, v. 2, n. 1, p. 27-32, 1995.

FRANTZIY, K.; ANANIADOUY, S.; MIMAZ, Hideki. **Automatic recognition of multi-word terms**: the C-value/NC-value method. Center for Computational Linguistics, Manchester and Dept. of Information Science, University of Tokyo, 2000.

GROOTJEN F.A. **Conceptual query expansion**. Faculty of Science, Mathematics and Computer Science, Radboud University Nijmegen. The Netherlands, 2006.

HOOPER, R.; PAICE C. **The lancaster stemming algorithm**. 2005. Disponível em: <http://www.comp.lancs.ac.uk/computing/research/stemming/index.htm>. Acesso em: 06 maio 2009.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, vol. 28, no. 1, pp 11-21, 1972.

JUSTESON, J.; KATZ, S. Technical terminology: some linguistic properties and an algorithm for identification in text. **Journal of Linguistic Engineering**, Cambridge, v.1, n.1,p. 9-27, 1995.

KROVETZ, R. Viewing morphology as an inference process. **International ACM SIGIR conference on Research and development in information retrieval**, Pittsburgh, Pennsylvania, p. 192-202, 1993.

LOVINS, J. B. Development of a Stemming Algorithm. **Mechanical Translation and Computational Linguistics**, v. 11 n. 1/2, p. 22-31, 1968.

MANDALA, R.; TOKUNAGA T.; TANAKA, H. **Combining multiple evidence from different types of theresaurus for query expansion**. Department of Computer Science. Tokyo Institute of Technology, Tokyo, 1999.

MILIOS E.; ZINCIR-HEYWOOD N.; ZHANG Y. Narrative text classification for automatic key phrase extraction in web document corpora. **Proceedings of the 7th annual ACM international workshop on Web information and data management**, Gremen, Germany, 2005.

ORENGO, V. M.; HUYCK, C. A Stemming Algorithm for Portuguese Language. **Proceedings...** of Eigth Symposium of String Processing and Information Retrieval. Chile, 2001.

PANTEL, P. ; LIN, D. **A Statistical Corpus-Based Term Extractor**. Department of Computing Science. University of Alberta, Canada, 2006.

PONCHIROLLI, Osmar. **O capital humano como elemento estratégico na economia da sociedade do conhecimento sob a perspectiva da teoria do agir comunicativo**. 2000. 95 f. Dissertação (Mestrado) - Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2000.

PORTER, M.F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p.130-137, 1980. Disponível em: <http://www.tartarus.org/~martin/PorterStemmer/def.txt>. Acesso em: 06 maio 2009.

RACHEL, T-W. L. B. H.L. O. **Automatically Building a Stopword List for an Information Retrieval System**. Department of Computing Science. University of Glasgow. UK, 2006

RODRIGUEZ y RODRIGUEZ, M. V. **Gestão do conhecimento**: reinventando a empresa para uma sociedade baseada em valores intangíveis. Rio de Janeiro: IVPI Press, 2001.

SIGHAL, A. Modern information retrieval: a brief overview. **Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**, p.1-9, 2001.

WIKIPEDIA. Disponível em:< http://pt.wikipedia.org/wiki/Copa_do_Mundo. 2005>. Acesso em: 06 maio 2009.

## SOBRE OS AUTORES

**Raphael Winckler de Bettio**

Bacharel em Ciências da Computação formado pela FURB (Blumenau). Mestre em Engenharia da Produção e Doutor em Engenharia e Gestão do Conhecimento. pela Universidade Federal de Santa Catarina Atualmente atua como consultor em Tecnologia da Informação na empresa Summa Technologies do Brasil.

E-mail: raphaelwb@gmail.com

**Alejandro Martins Rodriguez**

Nasceu em Montevidéu, Uruguai, em 28 de abril de 1963. Sua graduação foi em Engenharia Industrial Mecânica. Seu mestrado é em Engenharia de Produção pela Universidade Federal de Santa Catarina. Seu doutorado é em Engenharia de Produção pela Universidade Federal de Santa Catarina. Atualmente seu vínculo principal é no Instituto Superior Tupy, IST-SOCIESC, aonde atua como Professor dos Cursos de Engenharia Mecânica e Engenharia de Produção Mecânica; também, atua como Professor Colaborador no Programa de Pós-Graduação em Engenharia e Gestão de Conhecimento da Universidade Federal de Santa Catarina.

E-mail: aljmartins@gmail.com