

Linguarudo - Uma Arquitetura Lingüisticamente motivada para Recuperação de Informação de textos em português*

Rachel Virgínia Xavier Aires

Orientação:

Prof^a. Dr^a. Sandra Maria Aluísio
Dr^a. Diana Santos

Monografia apresentada ao Instituto de Ciências Matemáticas de São Carlos - USP, para o Exame de Qualificação, como parte dos requisitos para a obtenção do título de Doutor em Ciências - Área de Ciências de Computação e Matemática Computacional.

USP - São Carlos
Março de 2003

* Trabalho realizado com o auxílio da FCCN desde setembro de 2001.

ÍNDICE

1. INTRODUÇÃO	1
<i>1.1 Contextualização</i>	<i>1</i>
<i>1.2 Motivação e Relevância.....</i>	<i>3</i>
<i>1.3 Objetivos.....</i>	<i>5</i>
<i>1.4 Organização da Monografia</i>	<i>5</i>
2. RECUPERAÇÃO DE INFORMAÇÃO (RI).....	7
2.1 Processo de Recuperação de Informação.....	8
2.1.1 Linguagem de consulta	11
2.1.2 Técnicas de indexação.....	13
2.1.3 Modelos de Recuperação	16
2.1.3.1 Modelo Booleano.....	16
2.1.3.2 Modelo Vetorial	16
2.1.3.3 Modelo Probabilístico.....	18
2.2 RI: uma história	21
3. RI E PROCESSAMENTO DE LINGUAGEM NATURAL - INDO ALÉM DA FREQUÊNCIA DAS PALAVRAS	27
3.1 Índices	28
3.2 Interpretação das Consultas e Retroalimentação	31
3.3. Comparação entre documento e consulta	34
3.3.1 Segmentação de textos	34
3.3.2 Uso de citações e características estilísticas de um texto	35
3.4 Apresentação dos resultados e Diálogo	35
3.5 RI e PLN - Aplicações e Considerações.....	37
4. AVALIAÇÃO DE SISTEMAS DE RI	40
4.1 Abordagens para a avaliação	40
4.2 Relevância.....	41
4.3 Revocação, precisão e outras medidas	43
4.3.1 Medidas com preocupações atuais	46
4.4 O conjunto de teste	48
5. PROPOSTA E PLANO DE TRABALHO	51
5.1 Atividades Planejadas.....	54
5.2 Cronograma.....	61
5.4 Contribuições	62
5.5 Viabilidade e Recursos Disponíveis	62
BIBLIOGRAFIA E REFERÊNCIAS.....	64
GLOSSÁRIO	77

Lista de Figuras

FIGURA 1 - PROCESSO TÍPICO DE RI (BELEW, 2000)	8
FIGURA 2 - LEI DE ZIPF (ADAPTADA DE VAN RIJSBERGEN, 1979, P.16, FIGURA 2.1).....	9
FIGURA 3 - EXEMPLO DE ARQUIVO OU ÍNDICE INVERTIDO	14
FIGURA 4 - EXEMPLO DE ARQUIVOS DE ASSINATURA	15
FIGURA 5- EXEMPLO DE ÁRVORE DE SUFIXOS.....	15
FIGURA 6 - SIMILARIDADE DE DOCUMENTOS NO MODELO VETORIAL.....	17
FIGURA 7 - LINGUARUDO _ UMA ARQUITETURA LINGÜISTICAMENTE MOTIVADA PARA RI PARA PORTUGUÊS	53

Lista de Tabelas

TABELA 1- EXEMPLOS DE TÉCNICAS DA RI ADOTADAS POR FERRAMENTAS DE BUSCA.....	24
TABELA 2 - TÉCNICAS, RECURSOS E PESQUISAS QUE PODEM MELHORAR A QUALIDADE DOS SISTEMAS DE RI	37
TABELA 3 - EXEMPLOS DE PESQUISAS DE PLN PARA PORTUGUÊS QUE PODERIAM SER UTILIZADAS NA RI.....	38
TABELA 4- PONTUAÇÃO EM JULGAMENTO DE RELEVÂNCIA, PROPOSTA POR GWIZDKA & CHIGNELL (1999)	48

Resumo

Recuperação de informação (RI) textual é tema de pesquisas desde a década de 40 e vem crescendo junto com a Computação, utilizando os recursos disponíveis de várias de suas subáreas em cada época. Ao longo destes mais de 50 anos de vida, assumiu novas preocupações como a necessidade de tratar dados não estruturados e Recuperação de Informação multilingüe (*cross-language Information Retrieval*) presentes em ambientes cooperativos e na Internet e teve outras preocupações reforçadas no cotidiano, como encontrar técnicas de indexação mais rápidas e de recuperação com maior precisão. Recursos e técnicas de Processamento de Língua Natural (PLN) têm sido utilizados para melhorar o desempenho de algumas aplicações de RI como filtragem, categorização e busca, em diferentes momentos no processo de recuperar informação. O objetivo deste trabalho de doutorado é propor uma arquitetura para Recuperação de Informação Textual para português, que utilize PLN em todo seu fluxo – processamento de consultas, processamento de documentos e apresentação dos resultados. Para tanto, pretende-se estudar mecanismos e fenômenos lingüísticos em textos em português, para identificar recursos e técnicas de PLN que possam ser utilizados na identificação do “tema-assunto” de documentos e consultas. Proposta esta arquitetura, pretende-se desenvolver um protótipo de ferramenta de busca. A tarefa de busca foi escolhida por três motivos: 1) dentre os serviços disponíveis hoje na Internet, é o mais utilizado; 2) ainda frustra o usuário com a recuperação de muitos documentos irrelevantes ou de baixa qualidade; 3) engloba os três passos da Recuperação de Informação, possibilitando testar a idéia do uso de PLN em todo seu fluxo.

1. Introdução

A weekday edition of The New York Times contains more information than the average person was likely to come across in a lifetime in 17th-century England. – Wurman (1989)

1.1 Contextualização

Nos últimos anos, houve um crescimento explosivo do volume de informação¹. Livros, filmes, notícias, anúncios, música e, em particular, informação on-line surgem a todo o momento. Um estudo realizado na Universidade da Califórnia em Berkeley em 2000 (Lyman et al., 2000) sobre o volume de informação produzido anualmente no mundo em diferentes mídias estima que a produção mundial anual de conteúdo impresso, em filmes, óptico e magnético requereria cerca de 1.5 bilhões de gigabytes para ser armazenada. E que estes 1.5 bilhões seriam o equivalente a 250 megabytes por pessoa, isto é, para cada homem, mulher e criança na Terra. Segundo estatísticas da revista Inc. Magazine (1999), as pessoas gastam 150 horas durante o ano procurando por informação perdida e 45% das pessoas assistem TV e usam computadores simultaneamente. Especificamente sobre informações na Internet, o relatório da Universidade da Califórnia estima ser de aproximadamente 2.5 bilhões de documentos na Web, com uma taxa de crescimento de 7.3 milhões de páginas por dia, o que equivale a um valor entre 25 e 50 terabytes de informação, do qual de 10 a 20 terabytes seriam informação textual. E, considerando todo o tipo de informações disponíveis, incluindo a chamada web escondida (*deep Web*), são 550 bilhões de documentos interligados através da Web, sendo 95% desta informação publicamente acessível. O estudo traz ainda uma estimativa do volume de informação que circula por e-mail, listas de e-mail, usenet, ftp, IRC (*Internet Relay Chat*), serviços de mensagem e telnet. Apesar das dificuldades de estimar o fluxo de informação entre esses meios e os próprios autores terem dito que não estão considerando todos os dados, o volume impressiona são 748,412 terabytes de informação, somando e-mails, listas de e-mails, usenet e ftp.

¹ Os termos que estiverem sublinhados estão definidos no glossário.

Um ponto interessante de ressaltar sobre a Internet é o fato dela ser uma fonte de informação que permite que a mesma informação seja utilizada por várias pessoas como acontece com o rádio e a TV, ao contrário de outras mídias como livros e jornais em que cada exemplar, em geral, é lido apenas por uma ou duas pessoas (Lyman et al., 2000). A Internet mostra sua importância como mídia principalmente por uma característica atualmente crítica em nossa sociedade: a velocidade de mudança. A todo tempo acontecem inovações científicas, tecnológicas, culturais e sociais. Pesquisadores, educadores e pessoas de negócio frequentemente se sentem ultrapassados quanto a algumas mudanças no domínio em que trabalham. E mesmo enquanto pessoas comuns e não como profissionais, constantemente precisamos atualizar nosso conhecimento para nos adaptarmos às mudanças. Ou seja, estamos rodeados de informação e ao mesmo tempo sentindo que precisamos de mais. Por ser uma mídia atualizada a cada segundo por diversas pessoas, a Internet nos propicia sempre informações novas e atualizadas.

Tanta informação eletrônica nos traz também problemas. Há 20 anos, as pessoas contavam com processos relativamente simples de filtragem feita por editores de jornais, que selecionavam os artigos que seus leitores gostariam de ler, e pelas livrarias, que decidiam que livros expor, por exemplo. Hoje, este tipo de barreira para informações inúteis ainda existe, mas não é mais tão eficiente. Atualmente, as pessoas lidam com esta overdose de informação com esforço próprio, dicas de amigos e colegas de trabalho e um pouco de sorte. Desperdiçamos um grande número de horas procurando por informações que não sabemos onde estão armazenadas, tentando nos atualizar e lendo informações que nunca serão utilizadas por nós. Todo este esforço para gerenciar a informação acaba gerando custos extras com armazenamento de informação e pessoal para as organizações. Por exemplo, se um executivo ganha 60 mil dólares anualmente, 25 mil deste total são pagos apenas para que ele leia (Inc. Magazine, 1999). Além de gastos extras, uma consequência de lidar com um grande volume de informação são problemas para nossa saúde. Segundo psicólogos, lidar com tanta informação causa problemas psicológicos, físicos e sociais. O psicólogo David Lewis (1996) chegou a propor o termo “*Information Fatigue Syndrome*” para descrever os sintomas causados pelo excesso de informação, que incluem: ansiedade, capacidade pobre de decisão, dificuldades em memorizar e lembrar, e atenção reduzida.

Tantos problemas geraram um interesse maior pelo processo de gerenciar informação/conhecimento (*information management/knowledge management*). Gerenciar informação/conhecimento inclui: utilizar, buscar, armazenar, revisar, criar novo conhecimento ou atualizá-lo, ainda julgar, utilizar conhecimento externo e descartar conhecimento de pouca qualidade ou desatualizado. Apesar da gerência de informação/conhecimento ser objeto de estudo principalmente da área de administração e negócios, além de utilizar várias ferramentas de apoio da computação, está de alguma forma relacionada a áreas como descoberta de conhecimento (*knowledge discover - KD*), mineração de dados (*data mining*), mineração de textos (*text mining*), recuperação de informação (*information retrieval*), acesso à informação (*information access*), extração de informação (*information extraction*), pergunta e resposta (*question answering*), e filtragem de informação (*information filtering*). O enfoque maior destas áreas tem sido em métodos, modelos e técnicas para auxiliar a lidar com informação textual, principalmente a que está disponível na Web, devido ao grande volume de recursos e conhecimento e, seu maior alcance. Contudo, a pesquisa sobre técnicas para lidar com a sobrecarga de informação na Internet de forma a extrair o máximo de benefícios de seu conteúdo ainda está em seu início. Muito foi feito com relação a mecanismos de indexação, recuperação e navegação, mas pouco foi feito para garantir a qualidade da informação retornada, para garantir altos padrões de precisão (*precision*) e revocação (*recall*)².

1.2 Motivação e Relevância

Encontrar informação nesta nova mídia ou repositório de informação de tamanho monstruoso e pouca organização que é a Internet é uma tarefa difícil cuja importância tem aumentado consideravelmente de forma a poder ser considerada crítica. Desenvolver ou utilizar técnicas, métodos e modelos de Recuperação de Informação que garantam maior qualidade da informação retornada é uma tarefa essencial para ajudar qualquer usuário a lidar com a sobrecarga de informação, seja ele pesquisador ou alguém procurando por entretenimento. O conteúdo disponível na Internet duplica anualmente e as pessoas que incluem novos dados em sua maioria não sabem como

² Precisão e revocação são explicadas no Capítulo 4.

funcionam os sistemas para recuperação de informações. Com o aumento do conteúdo, vem também o aumento das fontes, que além de trazerem novos tipos de informação, têm causado o aumento do número de informações em outros idiomas que não o Inglês. Assim como nos primeiros anos da Internet o inglês ainda é o idioma predominante, mas não tanto como no princípio, atualmente o número dos usuários da Internet que são falantes nativos do inglês já se restringe a 50% (Lyman et al., 2000). Estima-se, que seja em torno de 5,090,230,228 palavras o tamanho da Internet em português (Aires & Santos, 2002).

Há bastante pesquisa em Recuperação de Informação no Brasil e em Portugal, são diversos profissionais trabalhando na Recuperação de Informação sob diferentes perspectivas como, por exemplo, psicólogos, bibliotecários, pesquisadores de interação usuário computador, pesquisadores de redes e pesquisadores de recuperação de informação. Entretanto, ainda há muito que ser feito para garantir a não exclusão de falantes do português da Sociedade da Informação. Esta não exclusão é um dos principais objetivos político-científicos do projeto da Linguateca³ (Processamento computacional do português), daí a necessidade de estudar a interação em português, desenvolver sistemas inteligentes que processem texto na rede em português e ajudem a encontrar informação em português, e avaliar o que existe para português e como melhorar seus padrões de qualidade.

Outro objetivo do projeto Linguateca é aumentar o uso das técnicas de PLN para o português - e uma das áreas óbvias, ainda que inexplorada, é a da RI. É, portanto, um dos sub-objetivos deste projeto identificar como e quanto será possível usar PLN e recursos sofisticados tais como corpora anotados em RI. O NILC⁴ (Núcleo Interinstitucional de Lingüística Computacional) compartilha deste objetivo. E tem também como objetivo verificar quais dos recursos de PLN já desenvolvidos por seus pesquisadores poderiam ser realmente aproveitados na tentativa de melhorar os resultados da recuperação de informação para português.

³ www.linguateca.pt

⁴ www.nilc.icmc.usp.br

1.3 Objetivos

Este trabalho se propõe a definir uma arquitetura para Recuperação de Informação Textual para português que resolva ou minimize consideravelmente o problema dos usuários de sistemas da RI na Web que é ter que lidar com um grande volume de documentos irrelevantes para ter acesso à informação procurada. A solução a ser investigada neste projeto, é explorar a fundo na arquitetura possibilidades de aplicação de técnicas e recursos de PLN não só na melhoria do mecanismo de comparação de consultas com o índice, mas também na melhoria da forma como os documentos são analisados e na análise do objetivo do usuário em uma dada consulta, ou seja, é tentar explorar a língua portuguesa em todas as fases do processo de RI – processamento de consultas, processamento de documentos e apresentação dos resultados.

Para tanto pretende-se estudar mecanismos e fenômenos lingüísticos em textos em português como relações semânticas e tópicos da área de estilística, para identificar recursos e técnicas de PLN que possam ser utilizados na identificação do “tema-assunto” de documentos e consultas. Proposta esta arquitetura, a mesma será utilizada pelo protótipo de uma ferramenta de auxílio à busca de informações armazenadas na internet.

Pretende-se, assim, rever a recuperação de informação sob a perspectiva de um profissional de PLN, desenvolvendo-se um trabalho extenso de levantamento de possibilidades para ajudar o usuário a escolher os itens que realmente deseja e precisa, poupando-o do esforço de lidar também com informações que não estava procurando. Pretende-se desenvolver um trabalho voltado para os falantes do português, que tenha possibilidades de levantar novas idéias para os profissionais que desenvolvem soluções inteligentes para a recuperação e gerência também para outros idiomas.

1.4 Organização da Monografia

Esta monografia está dividida em cinco capítulos. O Capítulo 2 define recuperação de informação, trata sobre o processo de recuperação de informação textual e apresenta

uma breve história sobre a evolução da recuperação de informação nos últimos 50 anos. O Capítulo 3 discute como os sistemas de recuperação de informação fazem uso de informações lingüísticas na tentativa de aumentar sua precisão e apresenta também possibilidades ainda não exploradas por sistemas comerciais. O Capítulo 4 descreve as principais formas de avaliação de sistemas de RI encontradas na literatura. O Capítulo 5 apresenta o plano detalhado de trabalho a ser desenvolvido neste projeto de doutorado.

2. Recuperação de Informação (RI)

*Data is like food. A good meal is served in reasonably-sized portions from several food groups.
It leaves you satisfied but not stuffed.
Likewise with information, we're best served when we can partake of reasonable, useful portions,
exercising discretion in what data we digest and how often we seek it out. - William Van Winkle*

Recuperação de informação (*Information Retrieval*) é a tarefa de encontrar itens de informação relevantes para uma determinada necessidade de informação expressa pela requisição de um usuário (consulta) e disponibilizá-los de uma forma adequada ao propósito da busca de informação do usuário. Por itens de informação entenda informação em diferentes mídias, tais como: textos, imagens (fotografias e mapas), vídeos etc. De acordo com a mídia tratada podemos classificar a RI como:

- RI textual ou RI documental (*Text Information Retrieval/Document Information Retrieval*);
- RI Visual que inclui RI de imagem e de vídeos (*Visual Information Retrieval*) (Ardizzone & La Casia, 1997);
- RI de áudio (*Audio Information Retrieval*) (Uitdenbogerd, 2000);
- RI multimídia (*Multimedia Information Retrieval*) (Chiaramella et al, 1996).

Em se tratando de RI textual esta ainda pode ser classificada como RI monolíngüe ou RI entre línguas (cross-language) (Oard, 1997, Peters, 2000). O que distingue um sistema de RI entre línguas de um sistema de RI monolíngüe é sua habilidade de recuperar documentos em uma língua natural diferente da utilizada na consulta. A RI entre línguas pode ainda ser classificada como bilíngüe ou multilíngüe.

A RI textual é o tipo de RI discutido neste capítulo. Apresentamos em detalhes o processo de recuperar informação na Seção 2.1, enfatizando a linguagem de consulta, a de indexação e modelos de recuperação. Concluimos o capítulo com uma breve discussão sobre a evolução das técnicas de RI desde a década de 40 até os dias atuais (Seção 2.2).

2.1 Processo de Recuperação de Informação

Dado um sistema de Recuperação de Informação como o da Figura 1, o processo de recuperar informação se dará em quatro etapas: i) representar cada documento em uma forma que possa ser “compreendida” pelo computador, ii) interpretar as consultas fornecidas, iii) comparar as consultas interpretadas com o conjunto de documentos indexados, e iv) apresentar os resultados de forma adequada ao propósito do usuário.

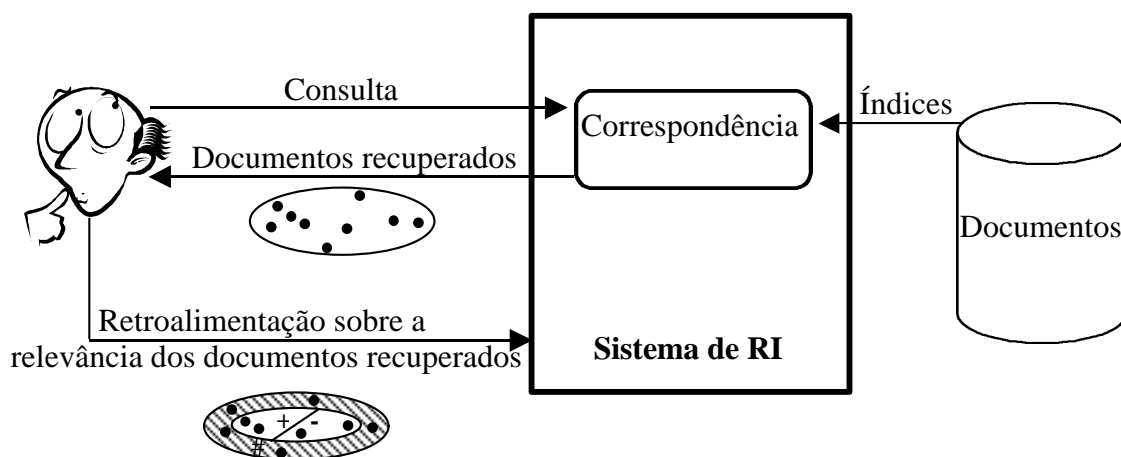


Figura 1 - Processo típico de RI (Belew, 2000)

A criação de **representações para os documentos** pode ser feita de forma manual ou automática. Para encontrar a forma de representação adequada pode ser analisado o conteúdo de todo o documento, apenas o resumo, alguns trechos ou até mesmo apenas uma lista de palavras. O resultado será uma lista de nomes, sendo que cada nome representa uma classe de palavras que aparece no texto de entrada. Um documento será indexado por uma classe se uma de suas palavras significantes for membro desta classe.

Luhn⁵ (1958) propõe que a frequência seja utilizada para extrair palavras e sentenças representativas de um documento. Dada uma frequência f de ocorrência e uma ordem r (*rank*) de sua frequência de ocorrência, então um gráfico relacionando f a r seria uma curva similar à mostrada na Figura 2, que diz que o produto da frequência de uso de uma palavra e sua ordem de importância é aproximadamente

constante. Luhn utiliza esta lei, lei de Zipf (1949), como uma hipótese nula para estipular dois pontos de corte, um inferior e um superior. As palavras que excedem o limiar superior são consideradas comuns e as abaixo do limiar inferior são consideradas muito raras.

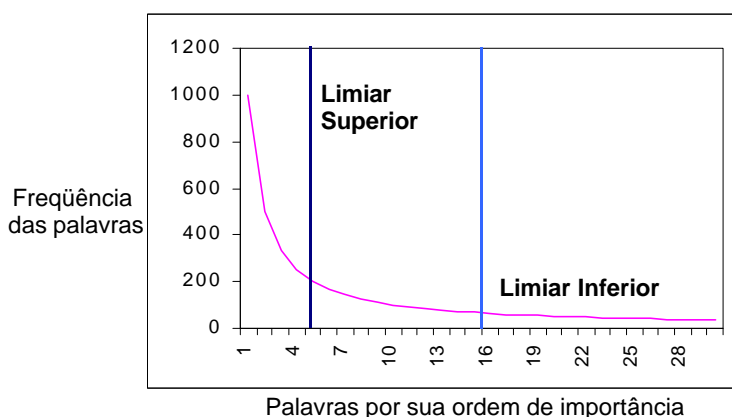


Figura 2 - Lei de Zipf (Adaptada de van Rijsbergen, 1979, p.16, Figura 2.1)

A remoção das palavras de alta frequência (*stop-words*) é uma forma de implementar o limiar superior – isto pode ser feito comparando a entrada com uma lista de *stop-words*. Um passo complementar seria remover sufixos (*suffix stripping*), assim muitas palavras equivalentes seriam mapeadas através de uma única forma. Outro passo seria checar os radicais, supondo que se duas palavras possuem o mesmo radical (*stem*) estas então se referem ao mesmo conceito e devem ser indexadas juntas. A saída final será um conjunto de classes, uma para cada radical detectado. O nome de uma classe é associado a um documento apenas se um de seus membros ocorre como uma palavra significativa no documento. A representação de um documento será então uma lista de nomes de classes, também chamada de índice de um documento ou palavras-chave (*keywords*). Caso a indexação seja realizada de forma probabilística, o resultado será um índice com pesos, assumindo-se assim que um documento pode ser sobre uma determinada palavra dado um determinado grau de probabilidade. A Seção 2.1.2 descreve as principais técnicas para construção de arquivos de indexação.

⁵ Hans Peter Luhn é considerado um dos precursores na Ciência da Informação e da RI.
<http://www.personal.kent.edu/~tfroehli/sighfis/luhn.htm>

Para aumentar a chance de se obter documentos relevantes, pode-se ainda contar com a ajuda de um tesauro (Yufeng & Croft, 1994), o que pode ser feito substituindo cada palavra-chave de um documento por cada uma das equivalentes.

Para estruturar a informação, os documentos podem ser agrupados de alguma forma que torne o processo de recuperação mais rápido. Isto pode ser feito através da clusterização (*clustering*) de palavras-chave ou da clusterização de documentos (Martin, 1995).

As **consultas** fornecidas como entrada serão interpretadas de formas diferentes de acordo com o tipo de consulta utilizado. Os tipos de consulta são apresentados na Seção 2.1.1. Para gerar uma consulta que possa ser analisada, o sistema pode utilizar também as técnicas de remoção de sufixos e checagem de radical, de tesauro e da associação de pesos aos termos da consulta — as mesmas estratégias citadas anteriormente quando falamos sobre geração de índices. Um tesauro pode ser utilizado: i) para substituir as palavras-chave de uma consulta quando a consulta original não retornou resultados ou retornou poucos e ii) na expansão da consulta que pode ser feita para se obter um número maior de resultados ou resultados mais precisos. A expansão de uma consulta pode ser feita gerando uma ou mais consultas através do uso de palavras sinônimas ou de palavras que têm alguma relação relevante com as que faziam parte da consulta original.

A **busca** em si dos documentos relevantes para uma consulta é feita comparando cada consulta aos documentos armazenados ou *profiles* contendo clusters de documentos. Para tanto um sistema adotará um Modelo de Recuperação ou adotará características de um ou mais modelos de recuperação. Um modelo de recuperação especifica quais são as representações utilizadas para documentos e necessidades de informação, e como estes são comparados (Turtle & Croft, 1990). Alguns exemplos de modelos são: Modelo Booleano (Paice, 1984), Modelo de Espaço Vetorial (Salton & McGill, 1983), Modelo Probabilístico (Maron & Kuhn, 1960), Modelos Booleano Estendido (*Extended Boolean models*) (Paice, 1984; Salton et al, 1983), Modelos de conjunto Fuzzy (*Fuzzy set models*) (Lee, 1995), Modelos Bayesianos (Ribeiro & Muntz, 1996) e Modelos Lingüísticos Estatísticos (*Statistical Language*

Models/Language Models) (Ponte & Croft, 1998). Na Seção 2.1.3 explicamos os modelos clássicos de recuperação.

A **saída** do sistema de IR costuma ser um conjunto de citações de documentos relevantes para uma dada consulta. As citações podem conter, por exemplo, título, nome de autores, trechos do texto que contêm os termos da consulta, data em que o documento foi publicado, há quanto tempo o documento está disponível no sistema, localização física ou eletrônica (web e intranets) do documento. Os resultados podem ou não estar ordenados segundo a relevância, já que alguns modelos de recuperação não permitem o cálculo de quão relevante é um documento. Os resultados podem também ser apresentados em grupos ou até mesmo em formas gráficas que explicam a relação entre os itens retornados como relevantes, como é o caso da meta ferramenta de busca Kartoo (www.kartoo.com).

Os resultados servem ainda como fonte de **retroalimentação** para o sistema, no caso de sistemas on-line em que é possível que o usuário mude sua consulta tentando melhorar o resultado da busca que está sendo realizada pelo sistema. Essa retroalimentação (*feedback*), pode acontecer através de mudanças feitas pelo próprio usuário diretamente nas consultas (sessões de consultas), pelo usuário fornecendo informações sobre sua satisfação ao sistema de forma explícita, ou automaticamente pelo sistema. O sistema pode tentar melhorar a qualidade dos resultados analisando os resultados que foram visualizados pelo usuário e em seguida modificar uma consulta acrescentando termos presentes nos resultados visitados ou gerando novas consultas com o uso de tesouros.

2.1.1 Linguagem de consulta

De acordo com Baeza-Yates & Ribeiro-Neto (1999), são três os tipos de consulta que pode ser formulada e submetida a um sistema de RI: palavras-chave (*keyword based query*), consultas por padrões (*Pattern-matching queries*) e consultas estruturais (*structural queries*).

As **consultas através de palavra-chave** são o tipo comumente aceito por sistemas de RI. Podem ser compostas somente por palavras soltas e, neste caso o

resultado retornado pelo sistema é um conjunto de documentos que contém pelo menos uma das palavras da consulta, ordenados pela frequência das palavras nos documentos (*term frequency*) ou pela frequência inversa (*inverse document frequency*). As consultas por palavras soltas podem ainda ser consideradas dentro de um contexto, procurando-se por uma frase consulta por frase, ou por palavras que estão a uma certa distância umas das outras consulta por proximidade (*proximity query*). As consultas por frase são na verdade uma seqüência de consultas por uma única palavra. As consultas por proximidade são uma forma mais flexível de consulta por frase, neste caso procura-se por uma determinada seqüência de palavras com uma distância máxima permitida entre elas. Esta distância pode ser medida em caracteres ou em palavras. As consultas por palavra-chave podem também ser compostas por palavras e operadores booleanos (consultas booleanas) ou podem ser formuladas em língua natural. No caso de consultas booleanas, um documento satisfaz ou não a consulta, não há como o documento satisfazer parcialmente a consulta.

Dizer que um sistema aceita consultas em língua natural, na maioria dos casos não significa que o sistema utilize sintaxe ou semântica para realmente interpretar o significado da consulta, isto em geral significa apenas que o sistema aceita que o usuário, ao invés de utilizar uma linguagem formal utilize língua natural. Ou seja, tais sistemas apenas extraem as palavras-chave de uma consulta para que esta seja representada para o sistema com várias palavras ou frases. Neste caso, qualquer documento que confira com parte da consulta é retornado como resposta, sendo que uma posição (*ranking*) melhor é associada aos documentos que conferem com o maior número de partes da consulta.

As **consultas por padrões** são utilizadas para permitir a recuperação de documentos com partes de texto que seguem propriedades pré-especificadas. Um padrão é um conjunto de propriedades morfológicas que precisa ocorrer em partes do texto, os tipos de padrão mais utilizados são: palavras, prefixos, sufixos, sub-cadeias de caracteres, intervalos (*ranges*), palavras semelhantes, expressões regulares e padrões estendidos.

Ranges são utilizados para cobrir quaisquer palavras que estejam entre um par de cadeias de caracteres seguindo a ordem alfabética, por exemplo, a range entre as

cadeias de caracteres retornar e rotular recupera cadeias de caracteres como retrair, retribuir, rigor e ritual. Já o padrão de palavras semelhantes permite encontrar palavras diferentes das fornecidas como entrada, procurando por pequenas diferenças (*error threshold*) causadas por erros de grafia ou de digitação. Por exemplo, a palavra retrair poderia ser encontrada a partir da entrada “retra ir”.

As expressões regulares são formadas por cadeias de caracteres e operadores como união, concatenação e repetição. Um exemplo é a consulta “pro (plem | teína) (a | s | ático) (0 | 1 | 2)*” que poderia encontrar palavras como “problema02”, “proteínas” e “problemático”. Os padrões estendidos são um subconjunto das expressões regulares com sintaxe mais simples. Podem fazer uso de classes de caracteres, expressões condicionais e caracteres coringa (*wild characters*). No caso das classes de caracteres, alguma posição no padrão irá conferir com um caractere de um conjunto pré-definido, por exemplo, alguns caracteres precisam ser dígitos e não letras. Uma expressão condicional indica que parte de um padrão pode ou não aparecer. A combinação permite encontrar qualquer sequência que confira, por exemplo, palavras que começam com “fo” e terminam com “ar”.

As **consultas estruturais** permitem que o usuário, além de utilizar características de conteúdo como fazia nas consultas por palavra-chave e por padrão, possa também utilizar características da estrutura do texto. As características a serem exploradas mudam de acordo com o tipo de estrutura seguida pelos textos: fixa, hipertexto ou hierárquica. Por exemplo, no caso de nossa caixa de entrada em um sistema de correio eletrônico, que é composta por e-mails, cada um com os campos: remetente, data, assunto e corpo de texto, é possível procurar pelos e-mails enviados por uma determinada pessoa com a palavra “avaliação” no campo assunto.

2.1.2 Técnicas de indexação

São três as principais técnicas para construção de arquivos de indexação (Baeza-Yates & Ribeiro-Neto, 1999): arquivos invertidos, arquivos de assinaturas e árvores de sufixo.

Arquivo Invertido é um mecanismo orientado por palavra baseado em listas de palavras-chave ordenadas, sendo que cada palavra-chave possui links para os documentos contendo aquela palavra-chave. Cada documento é associado a uma lista de palavras-chave ou de atributos, a lista é invertida e passa a não ser mais ordenada pela ordem de localização, mas sim por ordem alfabética. Cada palavra-chave ou atributo é associado a um peso. Após o processamento dos documentos esta lista é dividida em dois arquivos: de vocabulário e de endereçamento. O arquivo de vocabulário contém de todos os termos classificados e o arquivo de endereçamento contém uma série de listas, uma para cada entrada do arquivo de índices, cada uma com todos os identificadores dos documentos que contêm aquele determinado termo. Um exemplo é mostrado na Figura 3. O arquivo de vocabulário pode utilizar estruturas como vetores ordenados, estruturas hash e tries (*digital search trees*). A principal vantagem deste tipo de estrutura é sua facilidade de implementação e a principal desvantagem o alto custo para atualização do índice (Frakes & Baeza-Yates, 1992).

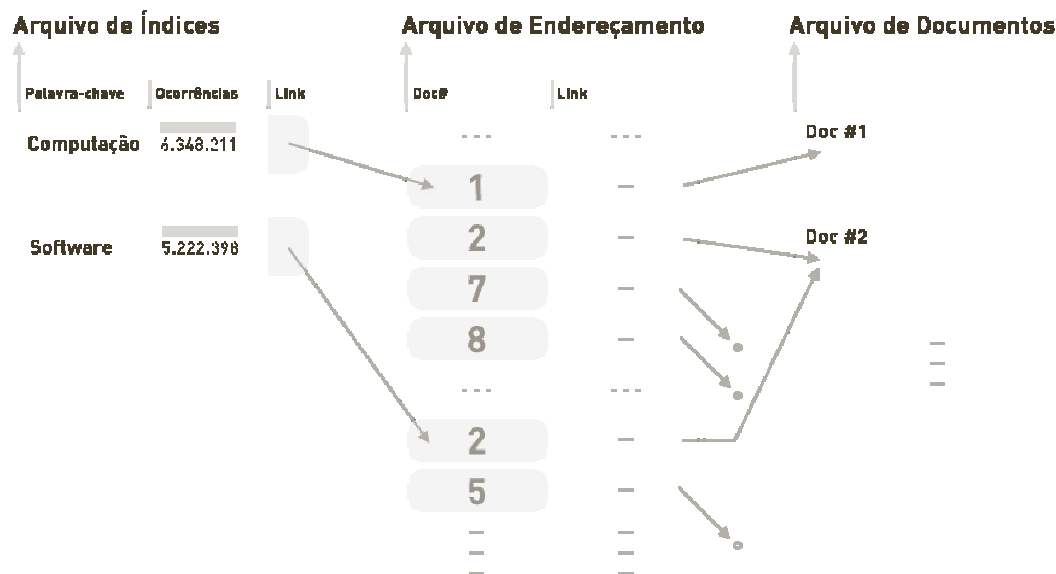


Figura 3 - Exemplo de Arquivo ou Índice Invertido

Arquivos de Assinatura são estruturas de indexação orientadas por palavra baseadas em hashing, são compostos por vários blocos de assinatura. As palavras são mapeadas para máscaras bit (*bit masks*) de B bits, que são a assinatura de cada

palavra, seu padrão de bits é obtido através de uma função hash. Os documentos são divididos em blocos lógicos contendo cada um, um número n de palavras. São apropriados para textos que não sejam muito longos, na maioria das aplicações os arquivos invertidos possuem uma performance superior aos arquivos de assinatura (Frakes & Baeza-Yates, 1992). Um exemplo pode ser visto na Figura 4.

Computer	0001	0110	0000	0110
Science	1001	0000	1110	0000
Graduate	1000	0101	0100	0010
Students	0000	0111	1000	0100
Study	0000	0110	0110	0100
Assinatura do blocco:	0001	0110	0000	0110

Figura 4 - Exemplo de Arquivos de Assinatura

No caso das **Árvores e Vetores de Sufixos** cada posição no texto é considerada como um sufixo. As árvores de sufixo são indicadas para consultas complexas, pois consultas frasais são caras de responder se utilizam arquivos invertidos, já para aplicações baseadas em palavras os arquivos invertidos têm melhor desempenho (Baeza-Yates & Ribeiro-Neto, 1999). Um exemplo pode ser visto na Figura 5.

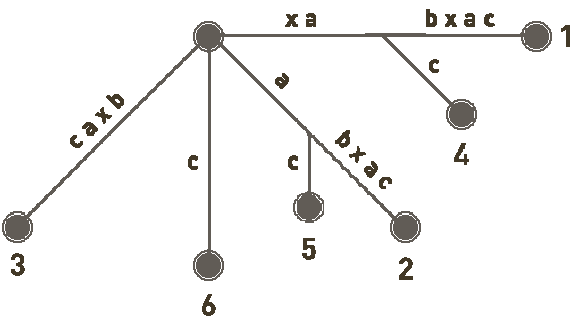


Figura 5- Exemplo de árvore de sufixos

2.1.3 Modelos de Recuperação

Um modelo de recuperação de informação prediz e explica o que um usuário irá considerar relevante dada sua consulta. São três os modelos clássicos seguidos por sistemas de RI para determinar a relevância de documentos: Booleano (Lógico), Vetorial, Probabilístico.

2.1.3.1 Modelo Booleano

O modelo booleano foi o primeiro modelo utilizado em RI e o modelo mais utilizado até o meio da década de 90 apesar das alternativas de modelo que surgiram desde o final dos anos 60.

O modelo Booleano considera uma consulta como uma expressão booleana convencional, que liga seus termos através de conectivos lógicos AND, OR e NOT. Neste modelo, um documento é considerado relevante ou irrelevante para uma consulta, não existe resultado parcial e não há informações que permitam a ordenação do resultado da consulta. O fato do modelo booleano não possibilitar a ordenação dos resultados por ordem de relevância é uma de suas principais desvantagens, já que esta classificação é uma característica considerada essencial em muitos dos sistemas de RI modernos, como por exemplo, nas máquinas de busca.

Outra característica deste modelo que pode ser considerada uma desvantagem no caso de usuários inexperientes é o uso de operadores booleanos. Para os usuários que conhecem bem álgebra booleana os operadores podem ser considerados como uma forma de controlar o sistema. Se o conjunto de resposta é muito pequeno ou muito grande, eles saberão que operadores utilizar para produzir um conjunto de respostas maior ou menor. No entanto, para usuários comuns os operadores booleanos não são intuitivos, pois seu uso é diferente do uso das palavras equivalentes a eles em língua natural. Por exemplo, se um usuário se interessa por música e por dança a consulta mais indicada seria “música OR dança” e não “música AND dança”.

2.1.3.2 Modelo Vetorial

No modelo de Espaço-vetorial, ou simplesmente modelo Vetorial, cada documento é representado por um vetor de termos e cada termo possui um peso associado que

indica seu grau de importância no documento. Em outras palavras, cada documento possui um vetor associado que é constituído por pares de elementos na forma $\{(palavra_1, peso_1), (palavra_2, peso_2), \dots, (palavra_n, peso_n)\}$.

Cada elemento do vetor de termos é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados em um espaço euclidiano de n dimensões (onde n é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso. As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos que possuem os mesmos termos acabam sendo colocados em uma mesma região do espaço e, em teoria, tratam de assuntos similares. Um exemplo é mostrado na Figura 6.

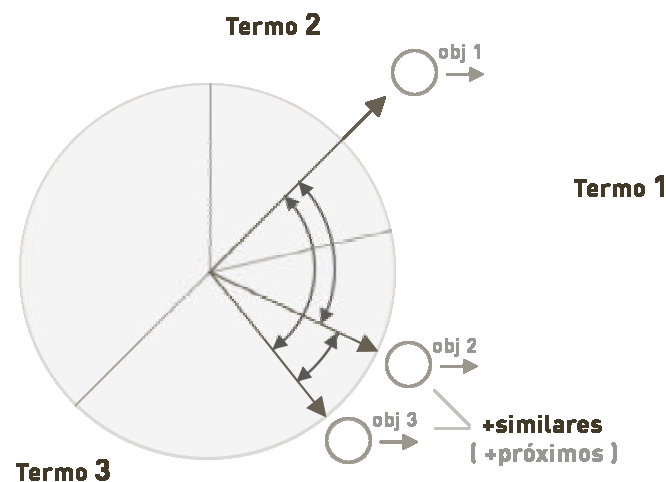


Figura 6 - Similaridade de documentos no modelo vetorial

Consultas também são representadas por vetores. Desta forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado. Os documentos mais similares (mais próximos no espaço) à consulta são considerados relevantes para o usuário e retornados como resposta para ela. Uma das formas de calcular a proximidade entre os vetores é testar o ângulo entre estes vetores. No modelo original, é utilizada a função cosseno (*cosine vector similarity*) que calcula o produto dos vetores de documentos através da fórmula:

$$\text{similaridade (Q,D)} = \frac{\sum_{k=1}^n w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^n (w_{qk})^2 \cdot \sum_{k=1}^n (w_{dk})^2}}$$

Dados:

- Q é o vetor de termos da consulta;
- D é o vetor de termos do documento;
- w_{qk} são os pesos dos termos da consulta;
- w_{dk} são os pesos dos termos do documento.

Calculados os graus de similaridade, é possível montar uma lista ordenada de todos os documentos ordenados por seus respectivos graus de relevância à consulta (*ranking*).

Uma desvantagem do modelo Vetorial é que não é possível incluir dependências entre termos no modelo, para modelar, por exemplo, frases ou termos que aparecem perto um do outro. Este modelo traz ainda duas dificuldades: a associação de pesos aos termos, que nem sempre é uma tarefa simples e, a implementação propriamente dita.

2.1.3.3 Modelo Probabilístico

No modelo probabilístico os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos. É baseado no princípio da ordenação probabilística (*Probability Ranking Principle*). Neste modelo, busca-se saber a probabilidade de um documento D ser ou não relevante para uma consulta Q_a . Tal informação é obtida assumindo-se que a distribuição de termos na coleção é capaz de informar a relevância provável para um documento qualquer da coleção. O modelo probabilístico é um dos poucos modelos que não necessita de algoritmos adicionais para associação de peso aos termos para ser implementado e os algoritmos de ordenação dos resultados são completamente derivados de sua teoria.

O modelo assume que a relevância de um documento é independente da relevância de todos os outros, e que um documento D será dito relevante para uma consulta Q_a quando: $P(+R_a/D) > P(-R_a/D)$.

Dados:

- R_a — documento é relevante para a consulta Q_a
- $-R_a$ — que o documento não é relevante para a consulta Q_a
- $P(+R_a/D)$ — probabilidade de que o documento D seja relevante para a consulta Q_a
- $P(-R_a/D)$ — probabilidade de que o documento D não seja relevante para a consulta Q_a

Assim, dada uma consulta Q_a , o modelo probabilístico atribui a cada documento D (como medida de similaridade) um peso W_{D/Q_a} , como sendo:

$$W_{D/Q_a} = \frac{P(+R_a/D)}{P(-R_a/D)}$$

Essa fórmula calcula a probabilidade de observação aleatória de D que pode ser tanto relevante quanto irrelevante. A teoria de Bayes auxilia a identificar para cada termo da consulta o grau de relevância e de irrelevância do documento. O valor final de probabilidade de relevância é dado pelo somatório dos graus de relevância de cada termo. Assim, aplicando a regra de Bayes:

$$W_{D/Q_a} = \frac{P(D/+R_a) \times P(+R_a)}{P(D/-R_a) \times P(-R_a)}$$

Onde:

- $P(D/+R_a)$ — probabilidade de que, dado um documento relevante para Q_a , este seja D
- $P(D/-R_a)$ — probabilidade de que, dado um documento não relevante para Q_a , este seja D
- $P(+R_a)$ — probabilidade de um documento ser relevante
- $P(-R_a)$ — probabilidade de um documento não ser relevante

Para calcular $P(D/+R_a)$ e $P(D/-R_a)$, como os termos indexados nos documentos são apenas presentes ou não presentes, o documento pode ser representado pelo vetor: $D = \{x_1, x_2, \dots, x_n\}$, $x_k \in \{0,1\}$. Ou seja, o peso para o termo

indexado x_1 pertence ao conjunto $\{0,1\}$. Colocando isso na fórmula, reescreve-se:

$$P(D/+R_a) = \prod_{k=1}^n P(x_k/+R_a)$$

Onde:

- $P(x_k/+R_a)$ é a probabilidade que evento descrito em x_k (presença ou ausência do termo k no documento D) ocorra, dado que o documento D é relevante para a consulta Q_a .

Ou seja, $r_{ak}=P(x_k=1/+R_a)$ é probabilidade de o termo k estar presente em D , sendo D relevante para a consulta Q_a .

$P(D/+R_a)$ pode ser reescrita da seguinte forma:

$$P(D/+R_a) = \prod_{k=1}^n r_{ak}^{x_k} (1-r_{ak})^{1-x_k}$$

Analogamente, $P(D/-R_a)$, probabilidade de o termo k estar presente em D , sendo D irrelevante para a consulta Q_a é dada por:

$$P(D/-R_a) = \prod_{k=1}^n s_{ak}^{x_k} (1-s_{ak})^{1-x_k}$$

•

Substituindo as duas últimas expressões na primeira (regra de Bayes) e considerando os logs, os pesos pode ser calculados da seguinte forma:

$$\begin{aligned} w_{D/Q_a} &= \sum_{k=1}^n x_k \times w_{ak} + C \\ x_k &\in \{0,1\} \\ w_{ak} &= \log \frac{r_{ak}}{1-r_{ak}} + \log \frac{1-s_{ak}}{s_{ak}} \\ C &= \log \frac{P(+R_a)}{P(-R_a)} + \sum_{k=1}^n \log \frac{1-r_{ak}}{1-s_{ak}} \end{aligned}$$

Para avaliar um documento é preciso simplesmente avaliar os pesos para os termos da consulta (w_{ak}), que também estão presentes nos documentos ($x_k=1$). A constante C que é a mesma para qualquer documento vai variar de consulta para consulta, mas pode ser interpretada como o valor de corte para a função de recuperação. A equação final pode ser escrita assim:

$$sim(D, Q_a) = W_{D/Q_a} = \sum_{k=1}^n x_k \times w_{ak}$$

W_{D/Q_a} é a medida de similaridade entre a consulta Q_a e o documento D . W_{ak} é o peso para o termo k na consulta, enquanto x_k é o peso para o termo k no documento. Uma vez que o valor de x_k é binário ($x_k \in \{0, 1\}$), pode se dizer que o modelo

probabilístico não atribui pesos aos termos nos documentos, ou seja, o modelo ordena os documentos apenas pela medida dos pesos dos termos da consulta (w_{ak}).

As duas principais desvantagens deste modelo são o fato de que para várias aplicações a distribuição dos termos entre documentos relevantes e irrelevantes não estará disponível e o fato de que o modelo define apenas uma ordenação parcial dos documentos.

2.2 RI: uma história

A importância de se ter uma coleção de informações científicas disponíveis para estudantes e pesquisadores vem sendo ressaltada há décadas por vários autores (Trivelpiece et al, 2000; Bowles, 1998), desde o trabalho pioneiro de Bush (1945). Foi ainda no final da década de 40 (Luhn, 1959; Ohlman, 1998) e durante a década de 50 que surgiram os primeiros trabalhos e sistemas de Recuperação de Informação (Lesk, 1995; Luhn, 1958). E foi em 1952 também que a expressão “*Information Retrieval*” começou a ser utilizada, após ser cunhada por Calvin N. Mooers⁶. Os sistemas desta primeira geração de sistemas de RI eram compostos basicamente por catálogo de cartões (Williams, 2002), contendo em geral nome do autor e título do documento.

A década de 60 foi uma época de muitos experimentos em RI. As métricas precisão e revocação (recall) usadas em processamento de sinais foram empregadas também para a RI. Surgiram as primeiras coleções para avaliação (Cleverdon, 1962) e foi também quando surgiu a idéia de retroalimentação sobre a relevância da busca (*relevance feedback*). Foi ainda na década de 60 que pesquisadores de Inteligência Artificial começaram a se questionar sobre os sistemas de RI se limitarem a encontrar documentos e os usuários ainda terem de lê-los para encontrar respostas a suas perguntas, momento em que começaram as pesquisas em sistemas de perguntas e respostas (*question-answering*). Foram várias as publicações na década de 60, relacionadas dentre outros tópicos a modelos probabilísticos, sistemas booleanos e modelo de espaço vetorial, por exemplo: Maron & Kuhns (1960), Becker & Hayes

⁶ <http://www.tracfoundation.org/mooers/mooers.htm>

(1963), Sparck Jones (1964), Salton (1968). Foi desenvolvido o sistema MEDLARS⁷ (Medical Literature Analysis and Retrieval System), o primeiro grande sistema de RI a utilizar uma base de dados informatizada e o processamento de consultas em *batch*.

Na década de 70 surgiram os primeiros processadores de texto e muitos textos começam a ficar disponíveis em formato eletrônico. São desenvolvidos os primeiros sistemas time-sharing – as consultas passam a ser apresentadas diretamente em terminais e o usuário pode ter a resposta imediatamente. Alguns exemplos são: NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT. É nesta década também que se intensificam as pesquisas em RI Probabilística. Alguns exemplos de publicações referentes a avanços teóricos e métodos de atribuição de pesos estatísticos da década de 70: Jardine & van Rijsbergen (1971), Salton (1975), Salton et al (1975a, 1975b), Van Rijsbergen (1979). Outro acontecimento importante da década de 70 se deu em 78 com a primeira conferência da Association for Computing Machinery (ACM) dedicada à Recuperação da Informação *Special Interest Group on Information Retrieval* (SIGIR⁸).

Na década de 80, o processamento de textos continua a crescer e o preço do espaço em disco começa a cair. Com isto e também com o desenvolvimento do CD-ROM muito mais informações (textos completos) ficam disponíveis, inclusive informações não textuais, o que faz despertar um interesse ainda maior pela RI multimídia. Porém, os avanços e novas direções da pesquisa nesta área, como por exemplo, a preocupação por técnicas de indexação capazes de lidar com grandes volumes de dados rapidamente, só são efetivamente vistos na década de 90. É na década de 80 que os sistemas de RI passam também a ser utilizados por não especialistas e que muitas das técnicas desenvolvidas anteriormente passam a ser realmente aplicadas.

Na década de 90, a comunidade de RI vê as tecnologias saírem da fase experimental para a fase de uso e sendo amplamente testadas devido à velocidade com a qual foram adotadas durante as décadas de 70 e 80 pelas aplicações comerciais.

⁷ Para uma descrição atualizada de MEDLARS, veja-se: Parris, Thomas M., 'Rx for Medical Information On-Line.' *Environment*, Vol. 40 No. 10, dezembro 1998, p. 3.
<http://environment.harvard.edu/guides/envbon/v40n10.html>.

É em 1992, por exemplo, que acontece pela primeira vez a conferência sobre Recuperação de Informação Textual *Text Retrieval Conference* (TREC⁹). Nos anos 90 passou-se a encontrar sistemas de RI com diversas finalidades: para bibliotecas comuns e digitais, específicos para serem utilizados por grupos de pesquisa com a finalidade de facilitar novas pesquisas e desenvolvimentos adicionais, sistemas associados a coleções de documentos e a ambientes computacionais de uma determinada instituição e, também sistemas amigáveis destinados a usuários com perfis diversos utilizando tipos diferentes de coleções de documentos em diferentes plataformas.

A RI começa também a gerenciar múltiplas coleções de documentos armazenadas em locais fisicamente dispersos, como por exemplo, estações de trabalho pessoais distribuídas como é o caso nas aplicações groupware, tendo, por exemplo, que localizar quais são as melhores bases de dados e mesclar os resultados destas buscas distribuídas. E também a se informar sobre soluções integradas, para que se possa integrar bem os sistemas de RI com os outros sistemas de uma organização.

É também no final dos anos 90 que surgem as máquinas de busca (*search engines*), diretórios (*directories*), e meta ferramentas de busca (*meta search engines*), adotando muitas características que até então haviam sido estudadas em RI, mas faziam parte apenas de sistemas experimentais, por exemplo: consultas em língua natural, resultados ordenados (*ranking*) e consultas através de exemplos. Isto fez com que a RI se deparasse com a necessidade de rever e melhorar as técnicas de indexação, de comparação dos índices com as consultas e as interfaces dos sistemas, devido aos tipos de dados encontrados na Web (distribuídos, voláteis, em grande volume, não estruturados, nem sempre de boa qualidade e heterogêneos), e ao perfil dos usuários destes sistemas. Na Tabela 1, mostramos como algumas das técnicas da RI discutidas nas seções anteriores foram adotadas por ferramentas de busca.

⁸ <http://www.acm.org/sigir/>

⁹ <http://trec.nist.gov/>

Tabela 1- Exemplos de técnicas da RI adotadas por ferramentas de busca

Características	Possibilidades Clássicas	Possibilidades adotadas pela maioria das ferramentas de Busca
Linguagem de Consulta	Consultas através de palavras-chave Consultas por padrões Consultas estruturais	Consultas através de palavras-chave
Linguagem de Indexação	Arquivos invertidos Arquivos de assinaturas Árvores de sufixo	Arquivos Invertidos
Modelo de Recuperação	Booleano Vetorial Probabilístico	Booleano Vetorial

Diferente dos sistemas de RI convencionais em que a coleção de documentos permanece relativamente estática, nos sistemas para a Web a alteração é constante, o que implicou na necessidade de técnicas mais eficientes para coleta de documentos na Web para que se pudesse cobrir uma grande porcentagem dela e se manter uma coleção atualizada (Belew, 2000). Além disto, estão disponíveis na Web documentos em vários formatos, por exemplo, SGML, HTML e pdf, o que fez com que os sistemas tivessem de ter uma espécie de analisador adicional para poder interpretar os diversos tipos de documentos (Belew, 2000). Outro problema que vem sendo pesquisado é como lidar de forma eficiente com o grande volume de dados, não só pensando-se no índice, mas também na ordenação dos resultados de forma mais eficiente para que o usuário não tenha que navegar por centenas de páginas (Plank, 2002). Outro ponto já bastante trabalhado é a qualidade questionável das informações disponíveis na Web, pois os sistemas da Web têm que lidar com *Spam* (Perkins, 2003), com páginas com conteúdo pornográfico e violento e com a dúvida sempre presente sobre a qualidade e origem das informações. A máquina de busca alltheweb (www.alltheweb.com) por exemplo, possibilita o uso de um filtro para omitir o que eles chamaram de conteúdo ofensivo e, a máquina de busca google (www.google.com) utiliza o algoritmo de premiação de páginas (*page ranking*) (Page et al, 1998) para tentar garantir que os usuários vejam primeiro os resultados de maior qualidade.

A outra principal diferença entre os sistemas convencionais e os sistemas na Web que causou mudança nas pesquisas nos anos 90 são os usuários. Na Web, não existe usuário típico, há usuários totalmente inexperientes e usuários com os mais diversos tipos de necessidades. São exemplos de usuários tanto uma criança que faz seu dever de casa, quanto uma secretária que procura o endereço da gráfica mais próxima e que deve entregar seu pedido em no máximo dois dias. Isto implicou em pesquisas sobre:

- (i) recuperação efetiva, que significa uma preocupação não só com precisão, mas em encontrar técnicas que não só funcionem bem para a maioria das consultas, mas que também não tornem difícil para o usuário se recuperar de erros graves ou pelo menos entender a origem dos erros;
- (ii) interfaces simples para usuários não especialistas e
- (iii) maiores esforços na interpretação de consultas, que podem: ser ambíguas, não terem indicado o objetivo, ou simplesmente conter um erro de digitação ou de ortografia.

Nos últimos anos, continua sendo dada atenção para os tópicos considerados na década de 90, e uma preocupação adicional da RI neste novo século é a busca de informações geográficas (Lopes & Rodrigues, 1996), considerando-se que o que está próximo do usuário é uma informação ainda mais relevante. As pesquisas nesta área são feitas tanto para serem aplicadas a sistemas da Web, quanto a sistemas tradicionais e mais recentemente também a sistemas para telefonia móvel (Loudon et al, 2002). No caso dos sistemas presentes em aparelhos celulares, fica ainda mais visível a importância deste tipo de pesquisa, já que os usuários deste tipo de sistema procuram informações para serem utilizadas logo após a consulta no local aonde se encontram, por exemplo, procurando pelo telefone da oficina mais próxima ao local aonde seu carro quebrou.

Nestes mais de 50 anos a RI foi crescendo junto com a Ciência da Computação. Utilizou os novos recursos disponíveis em cada época, por exemplo: novo hardware, novas técnicas de Inteligência Computacional e linguagens de representação de conhecimento, e também utilizou as pesquisas em Ciência da Informação, com o objetivo de encontrar os resultados mais relevantes para o usuário. Os avanços foram muitos, mas a RI continua se deparando com problemas que têm

sua origem na língua natural: muitos sinônimos, muitos significados, falta de habilidade dos usuários para expressar conceitos vagos que são importantes, erros de digitação e de ortografia (na linguagem escrita) e até mesmo indexação inconsistente. Problemas estes que se não tratados, ainda que parcialmente, impossibilitarão maiores avanços na melhoria da precisão dos sistemas de RI textual. No Capítulo 3 apresentamos alguns esforços que foram feitos neste sentido com uso de técnicas/recursos de PLN e lingüística e também apresentamos uma visão resumida do que o PLN ainda pode oferecer.

3. RI e Processamento de Linguagem Natural - Indo além da frequência das palavras

In interviews with experienced internet users we found that from the users' point of view, it is very likely that an information retrieval or filtering problem is framed as a problem of low quality of information, not of low topicality. – Karlgren (1999)

Como vimos no Capítulo 2, os sistemas de RI tradicionais se baseiam basicamente em estatística e matemática, considerando as palavras de um documento (e as da consulta) como unidades atômicas isoladas, utilizando primariamente algum uso da frequência destes termos no texto. Os métodos estatísticos aplicados a RI foram utilizados a partir da década de 60 após constatação da dificuldade em utilizar as técnicas de processamento de língua natural na RI (Bräscher, 1999). Entretanto, realizar a indexação de documentos e a busca utilizando somente métodos estatísticos compromete a eficácia, fato percebido desde o fim da década de 80 (Smeaton, 1990). Como veremos neste capítulo, as pesquisas em PLN voltaram a ser utilizadas na RI, durante a década de 90 e nos dias atuais. O que se busca é aumentar os níveis de eficácia da recuperação através do processamento da linguagem do texto (Smeaton, 1991 pg. 374 apud Bräscher, 1999) mesmo que o preço a ser pago seja o custo do processamento da língua natural. Entretanto, este custo impossibilita seu uso pervasivo em sistemas de RI cujo tempo de resposta é um dos critérios mais importantes.

Neste capítulo, apresentamos um resumo das técnicas que utilizam recursos que vão além da frequência das palavras ao analisar documentos e consultas para assim melhorar a precisão e/ou a revocação de sistemas de RI. Nas quatro primeiras subseções apresentamos algumas melhorias que poderiam ser feitas durante a criação de índices; interpretação de consultas e retroalimentação; comparação entre consultas e índice; apresentação de resultados e diálogo, respectivamente. Concluimos o capítulo na Seção 3.5 com algumas considerações sobre o uso de técnicas, recursos e pesquisas de PLN que podem ser utilizados na implantação das melhorias discutidas nas Seções 3.1, 3.2, 3.3 e 3.4.

É importante ressaltar que, neste texto, apesar de reconhecermos a importância da RI multilíngüe e da RI entre-línguas que faz com que os usuários de sistemas RI não tenham de rejeitar um resultado por este estar escrito em uma língua desconhecida ou diferente da utilizada na consulta, não apresentaremos as técnicas que são utilizadas nestes tipos de sistemas, por exemplo, os recursos de tradução e indexação semântica latente. Trataremos apenas da RI monolíngüe, que é o escopo deste trabalho.

3.1 Índices

Partindo do pressuposto de que os métodos estatísticos utilizam palavras, existem, pelo menos, duas formas de se incluir conhecimentos sobre a língua na indexação de documentos para melhorar sua performance: auxiliar na atribuição de múltiplos termos a um documento e auxiliar na combinação (conflation) de termos.

Considerar a frequência das palavras isoladamente é sem dúvida uma simplificação demasiado exagerada. Por exemplo, se desejamos indexar documentos com receitas de maria-mole, é mais intuitiva a combinação maria-mole¹⁰ do que representar os documentos por maria ou por representá-los através do termo mole. Ou seja, o uso de **múltiplos termos**, é uma forma de representar de maneira mais eficiente os tópicos tratados por um documento e de solucionar possíveis ambigüidades, através do relacionamento entre as palavras e seu papel dado aquele contexto de palavras. Por múltiplos termos, nos referimos a itens lexicais que co-ocorrem em um corpus, tanto termos que aparecem juntos como termos que aparecem próximos separados por um dado limite de palavras, ou seja, tanto palavras-compostas, nomes próprios e termos técnicos, quanto colocações (*collocations*). Os múltiplos termos cujo uso são mais defendidos na RI são os contituídos por sintagmas nominais (Arampatzis et al, 1998).

O uso de conhecimento lingüístico para selecionar múltiplos termos como representantes de um documento é assumido como uma forma de prover ganhos de precisão, por permitir distinções com granulidades mais finas entre termos similares,

¹⁰ Isto é, mesmo sabendo que maria-mole é uma palavra composta por justaposição, muitos sistemas ignoram este fato e trabalham com os dois componentes da palavra separadamente.

porém não idênticos, com estrutura interna diferente, ou estabelecendo relações mais elaboradas entre os elementos/termos identificados.

A forma mais simples de encontrar múltiplos termos é a análise de n-gramas para encontrar seqüências de palavras que são recorrentes em textos/documentos. Unindo informações estatísticas a informações léxicas, podemos extrair colocações (Smadja, 1993) em construções recorrentes, por exemplo: relações predicativas como as existentes entre verbos e objetos, sintagmas nominais idiomáticos e expressões frasais. Esta lista de múltiplos termos poderia ainda ser refinada através da análise da relação de dependência entre os termos, utilizando-se para isso frases sintaticamente analisadas. Esta segunda análise serviria para normalizar as entradas, por exemplo, análise por computador e análise computacional; ficha cadastral e ficha de cadastro; dor abdominal e dor no abdômen; queijo de minas e queijo mineiro.

Os termos técnicos são mais fáceis de serem encontrados devido ao fato de não serem facilmente modificados. Uma forma simplificada para sua extração é considerar que, caso um sintagma nominal tenha sido utilizado mais de uma vez na mesma forma dentro de um texto e entre textos, este sintagma seja um termo técnico. Note que nem sempre a freqüência evidencia um termo o que faz com que os sistemas estatísticos de extração de termos gerem muito silêncio, pois existem termos técnicos que aparecem com baixa freqüência em um corpus e não são encontrados, ou também gerem muitos ruídos, pois tais sistemas recuperam palavras com alta freqüência que não são termos e sim palavras sem valor especializado, mas que estão presentes nos textos como: estudo, tema, causa, processo, etc. (Estopà Bagot, 2001). Uma solução para os problemas de ruídos e silêncio dos sistemas de extração de termos é o uso de métodos híbridos, isto é, lingüísticos e estatísticos.

Combinar termos similares em um único termo de índice pode fornecer ganhos de revocação, permitindo que mais documentos com apenas diferenças triviais sejam identificados pelo mesmo conjunto de termos. A **combinação** pode ser feita olhando-se a morfologia, a sintaxe ou a semântica. No nível morfológico é feita através do uso de um *stemmer* ou remoção de sufixos, ou por palavras relacionadas em nível de paradigma morfológico, como, por exemplo: compromisso e comprometer-se; matado e morto; imprimido e impresso.

No nível sintático tenta-se agrupar sintagmas semanticamente equivalentes, mas sintaticamente diferentes, variações do tipo substantivo-sintagma preposicional e substantivo-adjetivo, por exemplo: divisão em frações e fracionamento; paralisia cerebral e paralisia do cérebro.

No nível semântico, a combinação é feita através do agrupamento de sinônimos, por exemplo: homicídio e assassinato; recuperação, recolha e obtenção. O agrupamento é feito tipicamente com o auxílio de tesouros, que são geralmente voltados para um domínio específico e são difíceis de padronizar, construir e manter. Alternativamente, pode ser feito uso de técnicas matemáticas, como a Indexação Semântica Latente (*Latent Semantic Indexing*) (Deerwester et al, 1990).

A indexação semântica latente tem seu funcionamento em torno da observação de que uma matriz de termos de índices por documentos é esparsa: a maioria dos termos não aparece na maioria dos documentos e a matriz conteria muitos valores nulos. Esta matriz pode ser então reduzida a uma matriz menor e mais densa, através de várias técnicas matemáticas. Os resultados de entrada são de alguma forma significados/sentidos de palavras e termos: agrupam termos que têm alguma relação entre si, o que presumivelmente seria útil em uma busca. O quanto se deseja reduzir a matriz é uma questão de quanta informação se está disposto a sacrificar para ganhar revocação originada pela combinação.

No caso de experimentos para inglês, o uso de múltiplos termos demonstrou um ganho mínimo que não justificaria o esforço de rodar e implementar métodos lingüísticos (Sparck Jones, 1999). A mesma afirmação é válida para os métodos de combinação, com exceção das técnicas utilizadas em *stemmers* e remoção de sufixos que são utilizadas em sistemas para o inglês com a justificativa de que a língua inglesa não possui características que justifiquem análise morfológica elaborada e que os custos de processamento gerados pela aplicação de *stemmers* ou pela remoção de sufixos são baixos. No entanto o inglês é uma língua tipologicamente diferente de várias línguas, inclusive o português e por isso os resultados e sugestões quanto a aplicações das técnicas variam. Para o português tanto o uso de múltiplos termos, quanto o uso de *stemmers* têm sido defendidos. Kuramoto (2002), por exemplo,

defende o uso de sintagmas nominais para representar documentos ao invés de palavras em seu estudo para português. Kuramoto enfatiza que os sintagmas nominais poderiam ser utilizados tanto através da simples substituição de índices que utilizam palavras em índices que utilizam sintagmas nominais, quanto no aproveitamento da organização hierárquica em árvores de sintagmas nominais, estrutura escolhida por ele em sua tese de doutoramento. Storb e Wazlawick (1998) defendem o uso de *stemmer* difuso para o português. No modelo proposto por eles, para cada par radical-sufixo é calculado um grau de certeza entre 0 e 1. Eles consideram que a semelhança entre significados de palavras, através da comparação dos radicais é determinada pelo reconhecimento correto de radicais e sufixos, por isto no cálculo do grau de certeza consideram a certeza para o radical e a certeza para o sufixo.

3.2 Interpretação das Consultas e Retroalimentação

Os modelos de RI da abordagem estatística tratam as consultas de forma semelhante a que tratam os documentos, ou seja, formam um vetor com as palavras da consulta extraindo as palavras muito freqüentes (*stopwords*), para posteriormente comparar este vetor com o vetor de palavras-chave dos documentos. Quando Luhn (1958) propôs seu modelo, a relação entre documentos e consultas não era a mesma com a qual nos deparamos atualmente – os documentos eram curtos, tipicamente resumos de textos e não textos inteiros. Atualmente, os documentos são textos que podem ser grandes, enquanto as consultas em geral são realizadas com apenas um pequeno número de palavras.

Tem sido observado que as consultas são geralmente pequenas, muitas com apenas duas ou três palavras (Abdulla et al, 1997; Cacheda & Viña, 2001a; Spink et al, 2002). Mecanismos baseados em análise lingüística muitas vezes precisam de mais material textual para modelar uma consulta e mesmo os mecanismos estatísticos provavelmente retornariam melhores resultados com mais informação já que este tipo de mecanismo é sensível ao volume de dados. Por isso, é interessante que os sistemas de RI possuam métodos para encorajar o usuário a produzir consultas maiores. As duas formas principais que poderiam ser utilizadas são: permitir que o usuário forneça sua consulta ao sistema em língua natural ou a expansão da consulta originalmente fornecida como entrada.

A opção de aceitar consultas em língua natural parte do pressuposto de que é mais fácil para um usuário inexperiente explicar suas necessidade de informação da forma com que geralmente faz em seu dia a dia. E que por isso o uso de perguntas conteria mais informação sobre a necessidade real do usuário do que o uso de palavras-chave. Apesar deste tipo de entrada ser aparentemente melhor para usuários pouco experientes, o tipo de usuário cada vez mais comum com os novos de sistemas de RI, ele não é o mais adequado para sistemas de RI em meios como telefones celulares, já que nestes sistemas é mais prática a digitação de entradas curtas. Além do que, mesmo as consultas em língua natural podem ser mal formuladas, por conterem erros de ortografia, por uma falta de clareza do que se deseja, ou por uso de palavras diferentes das utilizadas nos documentos, fazendo com que seja necessário a geração de novas consultas.

No caso da consulta ter sido mal formulada por conter erros ortográficos, o sistema pode procurar por palavras semelhantes às utilizadas na consulta para modificá-la automaticamente ou apresentá-las como alternativas para o usuário. Isto pode ser feito, por exemplo, utilizando *stemmers*, dicionários ou técnicas mais avançadas de correção ortográfica (veja, por exemplo, a máquina de busca Google: www.google.com).

Quando a consulta é uma consulta vaga ou ambígua que pode dar origem a muitos documentos irrelevantes, seu refinamento poderia ser feito através da adição de mais palavras - expansão da consulta. Um método para obter uma consulta mais refinada seria considerar a primeira consulta apenas como o início da busca e utilizar os resultados da primeira consulta neste processo de refinamento. Este processo pode ser feito com ou sem a interferência do usuário. Podem ser utilizados os documentos que o usuário tenha dito serem relevantes ou simplesmente considerar os primeiros documentos retornados pelo sistema como fonte para a extração de palavras para a próxima consulta. Considerar os primeiros documentos retornados parte da hipótese de que os resultados sejam relevantes, isto é, oriundos de um sistema com boa precisão, sendo que a consulta seguinte serviria então para aumentar a revocação. O sistema pode, então, automaticamente gerar novas consultas através das palavras extraídas dos documentos considerados relevantes ou julgados como relevantes e

então submeter estas consultas ao sistema, ou apresentar a lista de novas consultas ao usuário para que ele decida qual é ou quais são as consultas mais apropriadas. Analogamente, pode-se fazer uso dos documentos não relevantes na retroalimentação do sistema, os documentos irrelevantes são descartados na primeira interação e os termos presentes nestes documentos têm seus pesos diminuídos nas interações seguintes. Uma opção para a retroalimentação é que o sistema agrupe os documentos que julga semelhantes, para que o usuário selecione grupos de documentos semelhantes ao invés de documento por documento. Esta última opção é especialmente interessante se considerarmos o fato de que muitas vezes vários documentos contêm individualmente um pouco da informação que procuramos e que é o conjunto de documentos que atende totalmente a nossa consulta. Apesar das dúvidas dos pesquisadores quanto à eficiência da retroalimentação, estudos com usuários mostraram que estes aparentemente entendem como a retroalimentação funciona e encaram a retroalimentação como uma forma de ter mais facilmente controle sobre o sistema (Koenemann & Belking, 1996) sendo que a retroalimentação é utilizada atualmente em várias implementações (Robertson & Sparck Jones, 1996)

Uma outra opção é realizar a expansão das consultas utilizando tesauros. O uso de tesauros possibilita tanto a tentativa de gerar novas consultas não ambíguas, quanto uma forma de reformular as consultas que utilizam termos diferentes dos utilizados nos documentos. Os sistemas podem então incluir palavras nas consultas com a intenção de gerar consultas mais específicas e menos ambíguas, quanto substituir palavras por sinônimos ou variações para tentar encontrar documentos relevantes, mas que não contêm as mesmas palavras utilizadas na consulta. Um exemplo de utilização seria para as variações de uma língua. Por exemplo, em Portugal se utiliza mais a palavra investigação enquanto no Brasil se utiliza a palavra pesquisa. Exemplos de uso de tesauros na expansão de consultas em português são os trabalhos de Gonzalez e Lima (2001a, b, c). Entretanto, é importante ressaltar que os tesauros não são suficientes por si só como solução para tratar todos os tipos de ambigüidade, por exemplo, a hiponímia e a polissemia.

Para tratar mais casos de ambigüidade podemos utilizar outras técnicas do PLN, como mostrado no trabalho de Bräscher (2002). No entanto, apesar deste tratamento de ambigüidade ser bem sucedido em sistemas de PLN, sua aplicação de

forma eficiente na RI ainda é discutível e difícil. Por exemplo, a aplicação de técnicas de desambiguação de significado na RI não resultou em resultados mais precisos (Sanderson,1994).

3.3. Comparação entre documento e consulta

Muitas das máquinas de busca colocam mais esforço nas técnicas de indexação para evitar o uso de algoritmos complexos ou mais complicados na comparação entre consultas e índice para procurar os documentos relevantes. Considerando que os usuários dos sistemas de RI atuais já estão acostumados com resultados rápidos devido ao uso constante de sistemas on-line, a estratégia de utilizar algoritmos mais simples na comparação de índices e consultas é uma característica mais do que interessante para os sistemas atuais, não só para as ferramentas de busca. Por isto nesta seção tratamos de duas técnicas que aparentemente não aumentariam a complexidade do mecanismo de comparação: o uso de textos segmentados e as características do uso e estilo do texto.

Estas técnicas não estão diretamente relacionadas ao tópico de um documento, mas podem ser utilizadas para ajudar a encontrar quais são os documentos realmente relevantes para um determinado usuário. Razão disto é que mesmo que o documento trate do assunto procurado pelo usuário, ele pode ser tratado de forma mais profunda ou superficial do que o necessário.

3.3.1 Segmentação de textos

Muitas das abordagens estatísticas assumem que as palavras aparecem mais ou menos de forma aleatória em um texto, independentes umas das outras e das ocorrências anteriores. Quando na verdade, as palavras aparecem nos textos seguindo um padrão de distribuição governado pela progressão textual dos tópicos discutidos e convenções comunicativas (Katz, 1996). Se os segmentos de texto com mais chance de serem topicalmente pertinentes são escolhidos e os termos são pesados em comparação com os termos de outras sessões, este peso refletiria a aparência topical do texto melhor do que um modelo não-progressivo. Algumas técnicas úteis para isto são a sumarização e a segmentação. Uma abordagem deste tipo seria bastante útil já que sabemos que um

mesmo texto pode tratar de vários assuntos, dando um maior destaque para apenas alguns.

3.3.2 Uso de citações e características estilísticas de um texto

Documentos em geral têm referências para outros documentos, um exemplo de uso desta característica é a análise de citações para organizar material científico, já que eles têm citações explícitas e outros ponteiros para materiais similares. No caso da Web, podemos utilizar *hiperlinks* para julgar a informação de acordo com o material que aparece junto.

Além das referências outras características dos textos que poderiam ser facilmente exploradas são as variações estilísticas, que são tão freqüentes em textos sobre um mesmo assunto quanto a variação de assuntos entre textos do mesmo gênero ou variedade (Karlgrén, 1999). Estilo é a regularidade observável no discurso, é a repetição insistente de uma característica, a adoção continuada da mesma solução para contextos semelhantes. As variações de estilo podem acontecer em vários níveis, por exemplo, na escolha de vocabulário e de estrutura sintática, e estão relacionadas tanto com a audiência do texto quanto a outros fatores como as preferências do autor. Há algumas características que podem ser medidas e que dão uma indicação do estilo de um texto, por exemplo, a freqüência relativa de palavras longas e o tamanho das orações. Determinar o estilo de um documento pode auxiliar a determinar se aquele documento é interessante para um determinado usuário.

3.4 Apresentação dos resultados e Diálogo

A maioria dos sistemas de RI atuais não possibilita que o usuário entenda o contexto em que as informações estão inseridas, não suporta completamente a interação ou diálogo do usuário com o sistema, não auxilia na identificação rápida dos documentos relevantes e não consideram que usuários diferentes têm diferentes tipos de necessidade e comportamento de busca diferentes.

A maioria dos sistemas de RI atuais apresenta seus resultados no formato de uma lista, que algumas vezes é ordenada por relevância. Veja, por exemplo a máquina

de busca www.altavista.com.br. Este tipo de apresentação não permite que o usuário tenha uma visão ampla do contexto em que as informações estão inseridas e da relação entre os documentos, ou seja, não dão uma visão geral de como o sistema funciona, o que facilitaria o uso do sistema em suas novas consultas e até mesmo o refinamento da sua última consulta. Para informar o usuário sobre como o sistema funciona, podemos, por exemplo, ordenar os resultados agrupando-os pelas organizações ou autores que os produziram. O sistema pode ainda apresentar os resultados de forma gráfica utilizando-se das relações semânticas que o sistema interpretou a partir dos termos das consultas. Ou utilizar técnicas de classificação para agrupar os resultados de acordo com os sub-tópicos que os documentos tratam.

Geralmente, cada consulta é vista pelos sistemas como uma sessão de busca. No entanto, sabemos que muitas vezes os usuários aprendem como formular sua consulta ou até mesmo aprendem e desenvolvem mais sua visão sobre o tema procurado ao longo de várias interações, de várias buscas. Até porque, à medida que vêem os resultados têm uma visão melhor de que tipo de informação podem encontrar. Por isso, os sistemas de RI deveriam encarar a interação do usuário com o sistema como um diálogo, dando suporte a seqüências de consultas e não só a consultas individuais. Uma forma de fornecer este suporte seria disponibilizando em sua interface as formulações recentes de uma consulta e os resultados recuperados para cada uma, para que pudessem ser revistos durante o diálogo.

Como o número de resultados retornados pode ser muito grande, o sistema deveria auxiliar o usuário a julgar de forma mais rápida o que é realmente relevante. O que poderia ser feito tanto deixando clara a relação entre os resultados, quanto fornecendo o máximo de informações sobre o documento para ajudá-lo em seu julgamento de relevância, por exemplo, através de um sumário. E, quanto mais voltado para a consulta do usuário, mais o sistema facilita o julgamento da relevância. Tombros e Sanderson (1998) compararam o uso de sumários tradicionais, estáticos e predefinidos formados em geral pelo título e algumas das primeiras sentenças do documento, a sumários baseados na consulta (*query based summaries*) e concluíram que estes últimos melhoraram tanto a eficácia (*accuracy*) quanto a velocidade dos julgamentos de relevância dos usuários.

Cada usuário pode ter diferentes objetivos ao acessar um sistema de RI e, apesar de não podermos modelar todos os objetivos, podemos tentar identificar padrões de informação procurada e ter uma saída diferente para cada um destes padrões. Geralmente, os sistemas de RI possuem um formato fixo de saída independente do tipo de informação procurada. Isto poderia ser melhorado apresentando, por exemplo, antes da lista de resultados, a informação que se imagina é mais apropriada para aquele padrão ou até mesmo aquela consulta específica. Por exemplo, se a consulta é um nome próprio, apresentar antes da lista de usuários um resumo com a url de uma página pessoal, telefone e endereço. Ou se a consulta é formada pelo nome de uma instituição/organização, apresentar primeiro em destaque a url desta instituição/organização.

3.5 RI e PLN - Aplicações e Considerações

Levantando as melhorias que poderiam ser feitas em cada uma das fases da RI, encontramos vários autores defendendo a importância do uso de técnicas linguisticamente motivadas na RI, em especial o uso de técnicas e recursos de PLN em diferentes fases da RI. Os usos iriam desde análises mais profundas na criação de índices a análises superficiais na apresentação dos resultados. Vale ressaltar que apesar das muitas possibilidades não encontramos sistemas que fizessem uso simultâneo de várias técnicas linguisticamente motivadas em toda a RI. Por exemplo, Chandrasekar & Srinivas (1997) utilizaram informações sintáticas para eliminar resultados irrelevantes. Um resumo das técnicas, recursos e pesquisas poderiam ser utilizados para realizar as melhorias citadas nas quatro sessões anteriores é mostrado na Tabela 2.

Tabela 2 - Técnicas, recursos e pesquisas que podem melhorar a qualidade dos sistemas de RI

	Índice	Consulta	Comparação	Resultados	Retro-alimentação
Análise de tema de um texto	X				
Análise morfológica	X	X			
Análise sintática	X	X			
Classificação de textos quanto ao gênero			X	X	
Colocações	X	X			
Co-referência	X				

Estudos sobre a influência do tamanho dos documentos na qualidade do conteúdo	X		X	X	
Ontologias	X	X	X	X	
Pergunta-resposta				X	
Reconhecimento de nomes	X	X		X	
Segmentação de textos	X		X	X	
Sumarização	X			X	
Tesauros	X	X			X

Pareceu-nos ainda pela pesquisa bibliográfica que, apesar de termos muitas das técnicas e dos recursos de PLN utilizados por sistemas de RI para inglês e outras línguas disponíveis, também, para o português, nossas pesquisas de PLN não foram ainda aplicadas em toda sua potencialidade na RI. Na Tabela 3, mostramos exemplos de alguns dos estudos de PLN para português que poderiam ser utilizados na RI.

Tabela 3 - Exemplos de pesquisas de PLN para português que poderiam ser utilizadas na RI

Áreas	Referências
Análise morfosintática	Aires (2000)
Análise sintática	Bick (2000), Martins et al (2002)
Classificação automática	Ribeiro et al (1998)
Correção ortográfica	Silva (2001), Pelizzoni (2002)
Extração automática de relações semânticas	Gasperin (2001)
Extração de sintagmas nominais	Miorelli (2001), Vieira et al (2000)
Extração de termos múltiplos	Dias et al (1999), Dias e Nunes (2001)
Mapeamento de dependências sintáticas em relações semânticas	Gamallo et al (2002)
Pergunta e resposta	Lopes & Quaresma (1999), Cunha (1997)
Recuperação de informação geográfica	Padilha (1997)

Co-referência	Rocha (1999); Sant'Anna (2000)
Sumarização automática	Pardo et al (2002)

4. Avaliação de sistemas de RI

If information is power and riches, then it is not the amount that gives the value, but access at the right time and in the most suitable form¹¹.

Sistemas de Recuperação de Informação têm sido avaliados e comparados há vários anos, foi ainda na década de 60 que Cleverdon (1962) listou os seis critérios que segundo ele poderiam ser utilizados em uma avaliação: (i) cobertura da base de dados do sistema; (ii) tempo de resposta; (iii) revocação; (iv) precisão; (v) forma de apresentação dos resultados; (vi) esforço do usuário. Desde então revocação e precisão foram e continuam sendo os critérios mais utilizados, apesar de todas as discussões a respeito de suas deficiências e todas as medidas alternativas sugeridas (Gwizdka & Chignell, 1999). Precisão e revocação e a maioria das medidas alternativas sugeridas têm sua base no conceito de relevância, que é explicado na Seção 5.2. As avaliações que são baseadas em julgamentos de relevância são as chamadas **avaliações centradas no sistema**, outra alternativa são as **avaliações centradas no usuário**. Estas duas abordagens são descritas na Seção 4.1 aonde mostramos algumas de suas vantagens e desvantagens. Na seção 4.3 mostramos algumas das medidas baseadas em relevância mais utilizadas e também algumas das revisões feitas sobre estas medidas para que estas continuassem a ser utilizadas como métricas nos sistemas atuais. Na seção 4.4 apresentamos os detalhes que devem ser considerados na criação do conjunto de teste para uma avaliação baseada em relevância.

4.1 Abordagens para a avaliação

A RI pode ser avaliada segundo dois diferentes ângulos: sob o ponto de vista do sistema ou do usuário. As avaliações centradas no usuário analisam a interface dos sistemas e a interação do usuário com estas interfaces, são utilizadas para avaliar o comportamento (processo de explorar a informação), necessidades e satisfação dos usuários. Estas avaliações não seguem uma metodologia padrão de avaliação, fazem

¹¹ <http://www.dcs.shef.ac.uk/research/groups/nlp/extraction>

uso de técnicas e medidas de avaliação de outras áreas como, por exemplo, da área de Interação Homem-Computador (IHC) e da Psicologia Experimental. Os métodos utilizados são em geral qualitativos e incluem: entrevistas, observações, experimentos "*think-aloud*" e pesquisas para verificar a opinião do usuário sobre a informação recuperada.

As avaliações centradas no sistema analisam o desempenho técnico de um sistema e têm como foco principal verificar a eficácia, isto é sua capacidade de recuperar documentos relevantes e de não apresentar documentos irrelevantes como resposta a uma determinada consulta. O motivo da medição da eficácia ser o foco principal destas avaliações é a hipótese de que quanto mais eficaz for um sistema mais ele atenderá as necessidades do usuário. Diferentemente das avaliações centradas no usuário, que são feitas através de experimentos interativos, as centradas no sistema utilizam um conjunto de teste, que é composto por: coleção de documentos, lista de consultas/requisições e julgamentos de relevância. Exemplos de conjuntos de testes são os utilizados na avaliação conjunta TREC¹². Existem várias críticas às avaliações centradas no sistema: serem realizadas em ambientes de "laboratório" e não em ambientes reais; qual a credibilidade que se deve dar aos julgamentos de relevância já que este é um conceito subjetivo (Wu & Sonnenwald, 1999); e quão representativo é o conjunto de consultas e de documentos, uma vez que costumam ser voltados para o domínio da ciência e tecnologia.

Em contrapartida, este tipo de avaliação é bem mais barato e rápido do que as avaliações centradas no usuário, que podem durar de meses a anos e, também propicia a possibilidade de se comparar de forma prática e confiável diferentes sistemas ou diferentes versões de um mesmo sistema.

4.2 Relevância

A noção de relevância é **subjetiva**, diferentes usuários podem ter opiniões diferentes sobre a relevância ou não de um documento. Um usuário especialista em RI poderia, por exemplo, julgar um trabalho de graduação que encontrou como resposta a sua consulta "RI +web" como irrelevante já que não é uma revisão de um especialista da

área, já para um outro usuário aluno de graduação procurando por informação para o trabalho que irá entregar na manhã seguinte o resultado poderia ser considerado altamente relevante.

Além de subjetiva, a noção de relevância é **situacional**, o mesmo usuário pode fornecer julgamentos de relevância diferentes para os mesmos documentos e consultas. Por exemplo, uma professora quando preparando uma aula sobre RI na Web poderá considerar irrelevantes os artigos muito avançados sobre o tema, imaginando que seus alunos não irão compreendê-los já em outro momento procurando por artigos sobre o mesmo tema para se interar das novidades porque irá participar de uma banca de doutorado, os artigos avançados serão altamente relevantes. Ou seja, o julgamento de relevância depende da necessidade do usuário.

Relevância é também **cognitiva**. Além de se julgar como relevante ou irrelevante, se formos julgar diferentes graus de relevância acabamos por pensar nos resultados avaliados anteriormente. Se, por exemplo, o primeiro resultado é julgado como 10 em uma escala de 0 a 10 de níveis de relevância, o segundo como 8 e o terceiro como 6, é de se imaginar que o terceiro resultado é menos relevante que o segundo que é menos relevante que o primeiro, além disso, o terceiro documento é menos relevante que o segundo da mesma forma que o segundo é com relação ao primeiro. Tentar manter todos os julgamentos consistentes no caso deste exemplo seria uma tarefa árdua.

Relevância é também **dinâmica**. Uma consulta pode não ser totalmente satisfeita por um único resultado, muitas vezes a resposta procurada é encontrada através da união de vários resultados e de pequenas quantidades de informação encontradas em cada um deles. Podemos, também, ao começar uma busca não conhecer bastante o assunto e não reconhecer a relevância de um resultado em um primeiro momento, mas sim apenas após vermos referências para este primeiro resultado em um outro que consideramos relevante já em uma primeira avaliação. Podemos também começar uma busca com uma necessidade e, no final da avaliação dos resultados, por vermos que está não foi totalmente atendida, passar a considerar

¹² <http://trec.nist.gov/data.html>

os resultados que foram a princípio considerados irrelevantes como relevantes. Por exemplo, se minha consulta inicial é “presentes baratos e originais para o dia dos namorados” e encontro resultados que fazem referência a presentes originais, mas que não são baratos, estes resultados são irrelevantes, mas se chego ao final da lista sem encontrar nada barato posso mudar de idéia quanto à relevância dos resultados.

Além de a relevância ser dinâmica, cognitiva, situacional e subjetiva, em muitos casos os resultados não são um documento, se tem acesso apenas a partes do documento: título, primeiro parágrafo e citações bibliográficas. Neste caso o julgamento irá ainda depender de quão informativos são os dados fornecidos sobre o documento.

4.3 Revocação, precisão e outras medidas

Precisão e revocação são medidas baseadas na noção de documentos relevantes de acordo com uma determinada necessidade de informação. A revocação é utilizada para medir a habilidade do sistema de encontrar todos os documentos relevantes, já a precisão mede a habilidade de recuperar documentos que são em sua maioria relevantes.

$$\text{revocação} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes}}$$

$$\text{precisão} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos recuperados}}$$

Outra medida, embora não tão utilizada, relacionada aos julgamentos de relevância é a *fallout*. A *fallout* indica a proporção de documentos irrelevantes recuperados.

$$\text{Fallout} = \frac{\text{número de documentos irrelevantes recuperados}}{\text{número total de documentos irrelevantes}}$$

A relação entre estas três medidas pode ser medida através da fórmula abaixo, para os casos em que se conhece o parâmetro *G* (*generality*). *G* é a densidade de

documentos relevantes na coleção, número de documentos relevantes dividido pelo número de documentos que compõe a base do sistema.

$$Precisão = \frac{Revocação \cdot G}{(Revocação \cdot G) + Fallout \cdot (1 - G)}$$

Nas avaliações, o par de medidas precisão/revocação é o mais utilizado (van Rijsbergen, 1979; Belew 2000). Para cada consulta submetida ao sistema podemos calcular a precisão e revocação. Se a saída do sistema depende de um parâmetro como a posição em que o documento aparecia na lista de resultados ou o nível de coordenação (*coordination level*), pode ser calculado o par precisão/revocação para cada valor do parâmetro, por exemplo dada uma consulta com 3 termos, para os resultados com nível de coordenação 3, 2, 1 e 0. Dados os valores para o par precisão/revocação para cada valor do parâmetro pode-se construir a curva de precisão/revocação para cada consulta, como pode ser visto no exemplo mostrado no Gráfico 1. Para medir a performance geral do sistema, o conjunto de curvas, um para cada consulta é combinado de alguma forma para produzir uma curva média, como ilustrada no Gráfico 2.

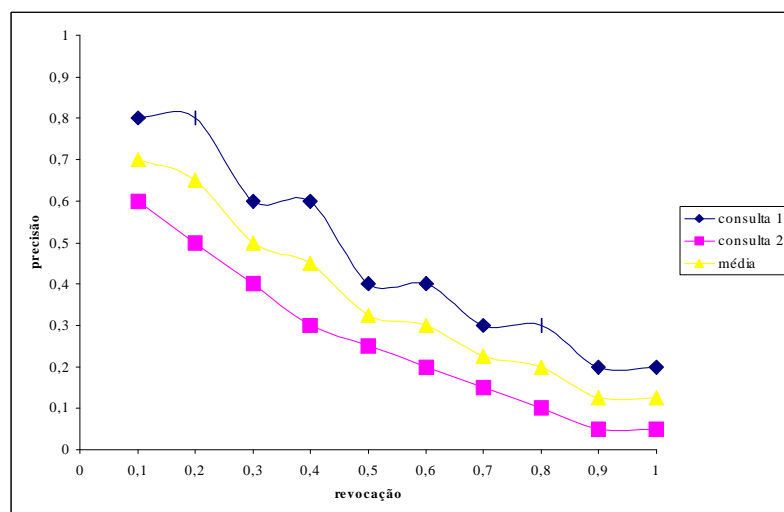


Gráfico 1 – Curvas de precisão/revocação

As curvas médias dos pares precisão/revocação para cada sistema podem ser calculadas para os valores de um determinado parâmetro, como o nível de coordenação, ou independentes de qualquer parâmetro. Neste último caso uma forma

de se montar a curva precisão-revocação do sistema é utilizando vários valores de revocação pré-estabelecidos. Neste caso são calculadas a média das precisões (precisão média) para cada curva para cada um dos valores de revocação preestabelecido.

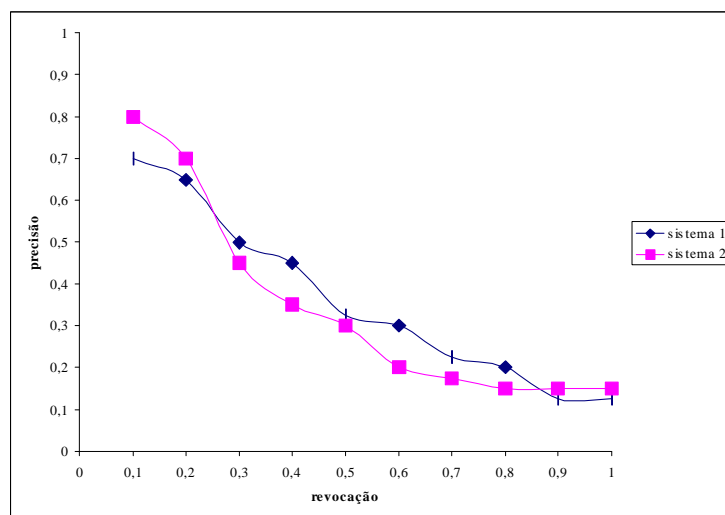


Gráfico 2 - Comparação de sistemas utilizando-se as curvas de precisão/revocação

Há sempre discussões sobre revocação e precisão serem ou não as medidas mais apropriadas para estimar eficácia. Algumas questões levantadas são: (1) precisão e revocação são mesmo medidas confiáveis?, (2) o quanto pequenas diferenças na revocação e precisão afetam o sucesso de uma busca?, (3) não seria mais interessante adotar uma medida que considerasse também o número de documentos irrelevantes?, (4) como medir a revocação se não existir documento relevante no conjunto de documentos?, (5) como medir a precisão se nenhum documento for recuperado? Além destas discussões o uso de pares de medidas deu origem a várias tentativas de criar medidas compostas, por exemplo, a medida F, a medida E, e a Diferença Simétrica Normalizada, apresentadas a seguir. A medida F associa revocação e precisão de uma forma que ambas têm de ser altas para que a medida tenha um valor alto. Já a medida E (*E-measure*) permite que se coloque ênfase na precisão ou na revocação. Quando b é 1 significa que revocação e precisão têm o mesmo peso e a medida E passa a ser igual à medida F. Quando se associa um $b > 1$ o peso da precisão é maior do que o da revocação e quando $b < 1$ o peso da revocação é o maior. A Diferença Simétrica Normalizada fornece a diferença proporcional entre o conjunto de documentos

relevantes e irrelevantes recuperados por um sistema. Quanto menor a diferença, melhor o sistema em recuperar todos os documentos relevantes para uma dada consulta. Uma discussão mais detalhada a respeito de medidas compostas pode ser encontrada em van Rijsbergen (1979) e Belew (2000).

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

$$E = \frac{(1+b^2)PR}{b^2P+R} = \frac{(1+b^2)}{\frac{b^2}{R} + \frac{1}{P}}$$

$$DiferençaSimétricaNormalizada = 1 - \frac{1}{\frac{1}{2}\left(\frac{1}{P}\right) + \frac{1}{2}\left(\frac{1}{R}\right)}$$

Encontramos ainda na literatura, também relacionada a julgamentos de relevância, a medida mediana (Greisdorf & Spink, 2001) e as medidas subjetivas: novidade (*novelty*) e cobertura (*coverage*). A novidade mede a habilidade do sistema de encontrar nova informação sobre um tópico, a proporção de documentos relevantes recuperados que o usuário não conhecia. A cobertura indica a proporção de documentos relevantes recuperados que o usuário já conhecia.

4.3.1 Medidas com preocupações atuais

As características desejáveis de uma medida de eficácia e seus comportamentos em diferentes casos são estudadas há anos e as medidas mais bem entendidas por serem as mais estudadas são precisão e revocação, ainda assim, a decisão de quais medidas utilizar depende da aplicação, e há sempre discussões sobre a confiabilidade de tais medidas (Su, 1998). No caso, por exemplo, dos sistemas na Web é de estranhar que as medidas revocação e precisão continuem como as medidas mais utilizadas, por dois motivos:

(1) Na Web não conhecemos o número total de documentos relevantes não selecionados como resposta pelo sistema, e como o número de respostas é, em geral, grande, é difícil obtermos também o número de documentos relevantes retornados como resposta e o número de documentos irrelevantes retornados como resposta;

(2) Além da escolha da melhor medida estar relacionada à aplicação, ela também está relacionada à situação, que não pode ser considerada em aplicações de uso generalizado como, por exemplo, nas máquinas de busca. Para este caso, em particular, em que os usuários são diversos, não é possível interpretar ao certo a relevância ou não de um julgamento. Por exemplo, um professor preparando o material a ser utilizado em suas aulas pode desejar encontrar com uma máquina de busca toda a bibliografia disponível sobre o assunto que será abordado; já para um aluno que está escrevendo um trabalho para o dia seguinte basta encontrar uma referência relevante. Para o professor deste exemplo a revocação é a medida mais importante, para o aluno é a precisão. No semestre seguinte este mesmo professor pode estar à procura de bibliografia sobre o mesmo assunto, mas neste segundo instante a medida que diz mais a respeito da qualidade do sistema para ele seria novidade. Ou seja, na Web precisão e revocação não são nem mesmo aceitáveis como medidas únicas. Para superar o obstáculo do grande número de documentos que deveria ser julgado, estas avaliações são feitas apenas sobre os primeiros resultados, e o número de resultados julgados varia de um estudo para outro, por exemplo, os 20 primeiros ou os 10 primeiros.

Outras características de sistemas na Web que já influenciam ou deveriam influenciar nas medidas quantitativas utilizadas, são os fatos de que para os usuários destes sistemas:

- (1) Em muitos casos, encontrar um único resultado relevante é suficiente.
- (3) Um documento que não responde completamente à pergunta, mas tem links para documentos que respondem pode ser considerado um documento relevante.
- (4) Muitos usuários não conferem todos os primeiros 10 ou 20 resultados quando não encontram nenhum documento relevante entre os primeiros listados.
- (5) Itens duplicados ou que não levam a nenhum documento não são relevantes.

Algumas avaliações já consideram algumas destas características. Alguns exemplos, de medidas atualizadas para serem utilizadas em avaliações de máquinas de busca, são as formas propostas por Gwizdka & Chignell (1999) para calcular a precisão: precisão total (*Full precision*), melhor precisão (*Best precision*), precisão útil (*Useful precision*) e precisão objetiva (*Objective precision*).

A precisão total considera a pontuação que foi associada a cada resultado segundo a Tabela 4.

Tabela 4- Pontuação em julgamento de relevância, proposta por Gwizdka & Chignell (1999)

Pontuação	Descrição
3	relevante
2	Parcialmente relevante ou contém um link para uma página de pontuação 3
1	Pouco relevante. Menciona rapidamente O tópico ou contém um link para uma página com pontuação 2
0	Não relevante ou link inválido

A melhor precisão considera apenas os hits mais relevantes, ou seja de pontuação 3. A precisão útil considera apenas os resultados com pontuação maior que 2, ou seja, os mais relevantes e os que contêm links para os mais relevantes. A precisão objetiva não requer julgamentos de relevância, é baseada na presença ou ausência dos termos requisitados e na distinção entre links bons e ruins (duplicados ou inválidos).

4.4 O conjunto de teste

São duas as formas utilizadas em geral para criar o conjunto de teste composto por documentos, consultas e julgamentos de relevância:

(1) Manual

Cada documento é julgado quanto à relevância em relação a cada consulta. Neste caso já existe uma base prévia de documentos, por exemplo, composta de notícias de jornais ou de artigos científicos. A lista de consulta em geral não é formada por consultas reais, mas sim por consultas elaboradas por bibliotecários, ou profissionais de RI que querem testar casos específicos que poderiam ser problemáticos ou mais difíceis para o sistema.

(2) Automática

Cada consulta é rodada em vários sistemas e os resultados são agrupados e apenas uma determinada proporção dos primeiros resultados é julgada. Neste caso a base de documentos será formada pelos resultados das consultas. A listas das consultas neste caso pode ter sido formada de consultas extraídas de uma lista de consultas reais, ou

pela elaboração de consultas por especialistas de RI, especialistas no domínio ou bibliotecários.

Faz parte também da estratégia de definição do conjunto de teste, no caso da criação ser manual, definir que tipos de documentos comporão a base de documentos, quem elaborará as consultas e em que número e como será feito o julgamento de relevância. No caso da criação automática deve ser definido de onde e como as consultas serão extraídas ou por quem e quantas serão elaboradas, como será feito o julgamento de relevância e quantos resultados de cada consulta comporão a base de documentos.

Os tipos de documentos que serão utilizados para compor a base são definidos pensando-se no que se deseja avaliar, por exemplo, um sistema para encontrar novidades na área médica. E no caso de se querer avaliar com uma base de documentos o mais genérica possível, o que se encontrar disponível. No caso de formar a base com documentos que são retornados como resultados das consultas, o número de resultados a ser aproveitado varia de um estudo para outro e depende da disponibilidade de juízes para efetuar posterior julgamento.

O número de consultas utilizadas em ambos os casos também não segue um padrão, há avaliações, por exemplo, que foram feitas com 3 (Pratt & Fagan, 2000; Notess, 2000), com 15 (Gwizdka & Chignell 1999; Notess, 1999) e com 50 consultas (Hawking et al, 1999). Mesmo quando as consultas que farão parte da lista são consultas reais é possível que se faça a opção de filtrar as consultas, para que se possa, por exemplo, remover as relacionadas com um determinado assunto, como pornografia, ou para que se possa remover as consultas que não possuem um objetivo claro. Em geral as consultas são julgadas apenas como relevantes ou irrelevantes, mas há alguns estudos que adotam níveis de relevância (Su et al, 1998; Gwidka & Chignell, 1999).

Antes de julgar a relevância é necessário que se defina o que é um documento relevante para cada uma das consultas. E também se um documento relevante é qualquer documento que trate do assunto ou se um documento é relevante apenas se responde por completo à consulta. É importante também que se instrua os juízes

dizendo como eles devem proceder, por exemplo: se não deve importar a veracidade da informação, se basta abordar o assunto para o documento ser relevante, se devem considerar ou não seu conhecimento prévio sobre o tema para determinar a relevância, se não devem deixar documentos julgados anteriormente interferir no julgamento do documento atual, etc.

5. Proposta e Plano de trabalho

"We need a new generation of Web searching tools based on a more thorough understanding of human information behaviours. Such tools would assist users with query construction and modification, spelling, and analytical problems that limit their ability or willingness to persist in finding the information they need." - Spink et al (2002)

A Lingüística, como ciência da linguagem, trata, dentre outros aspectos da língua, também de textos, discurso e diálogo. A Recuperação de Informação é em grande parte voltada para textos e seu conteúdo. Por isso, são vários os pontos em sistemas de RI tradicionais e comerciais em que resultados de pesquisa em lingüística poderiam ser aplicados, e também várias as oportunidades para que a RI forneça aos lingüistas matéria-prima para análises sobre o uso da língua. No entanto, nenhuma destas duas direções tem sido amplamente explorada. Uma justificativa para o pouco contato entre estas duas áreas talvez seja a forma diferente de analisar os documentos pelas duas áreas.

Os sistemas clássicos de RI vêem os documentos como seqüências de informação a respeito de um tópico, e assumem palavras e termos como indicadores de tópicos. Suas técnicas para análise e organização dos documentos são basicamente focadas nas palavras e em sua frequência em cada texto/documento analisados como um todo. Já, a lingüística acredita que as expressões da língua são formadas por palavras, que formam orações (cláusulas), que estão estruturadas em torno de um texto/discurso. Para a lingüística, palavras estão inseridas em uma situação, têm falante e estrutura de tópicos independentes e previsíveis que podem ser descritas formalmente.

Até o presente momento, como visto no Capítulo 3, ainda são poucos os recursos lingüísticos utilizados pelos sistemas de RI. Um exemplo do uso que se faz da análise morfológica é a tentativa de verificar variantes de uma palavra e a análise de palavras-compostas. Um dos usos mais elaborados da análise sintática é na interpretação das consultas e no agrupamento de variantes. E a análise semântica é feita apenas de forma implícita, quando os sistemas, além de se basearem na

frequência das palavras, levam em consideração a co-ocorrência de palavras. Outros trabalham também com léxicos, bases de conhecimento e redes de ontologia. O que não é uma afirmação de que não existem recursos mais avançados que tenham sido explorados pelo PLN, mas apenas que pouco do que foi explorado em PLN foi utilizado na RI. Acreditamos que isto se deva a três fatores: (i) poucos pesquisadores de RI são também pesquisadores de PLN e, apesar de possuírem uma visão empírica dos problemas que a língua natural causa para seus sistemas, não conhecem os esforços já realizados pelo PLN em resolvê-los através de conhecimentos lingüísticos; (ii) acredita-se que o custo-benefício de utilizar técnicas mais avançadas seria muito baixo, que o aumento da complexidade seria alto com um grande aumento no tempo de resposta e um pequeno ganho na precisão dos sistemas; (iii) assume-se que técnicas que tenham falhado para aumentar significativamente a precisão em sistemas para o inglês também falhariam em qualquer outra língua, o que pode ser um engano sério, já que o inglês é uma língua tipologicamente diferente, que depende mais da ordem das palavras do que muitas outras línguas e possui morfologia não tão rica quanto outras línguas, características que deveriam ser consideradas não só quando pensando em métodos lingüísticos, mas também no desempenho de métodos puramente estatísticos.

Acreditamos que a principal falha dos sistemas de RI atuais não é o pequeno número de recursos explorados, mas a forma como estes foram explorados. Pela revisão da literatura vimos que os sistemas atuais utilizam recursos de PLN em momentos isolados, por exemplo, para expansão de consultas ou na criação do índice. Em nosso levantamento bibliográfico não encontramos um sistema em uso que explorasse a língua com outras técnicas além das estatísticas em todas as etapas do sistema, isto é: para a criação de índice, interpretação da consulta e casamento entre consulta e documentos. Mesmo os sistemas que utilizam léxicos são frágeis, no sentido de que têm a mesma limitação que os modelos puramente estatísticos, pois eles modelam as unidades de significado/sentido como relativamente atômicas, não modelam relações, dependências, ações ou eventos coisas das quais o discurso depende. Temos de nos lembrar de que palavras são vagas, ambíguas, podem ser, por exemplo, tanto polissêmicas quanto incompletas, isto é, palavras podem ter vários significados e o mesmo objeto pode ser expresso através de várias palavras. Por isso precisamos utilizar em nossos sistemas modelos semânticos mais elaborados, que

modelem a língua em uso, que considerem a temporalidade e topicalidade dos termos, e também as relações e elementos gramaticais, o que significa que precisamos tentar entender e modelar o ciclo de vida dos termos na língua, o ciclo de vida das referências no discurso e a conexão entre ambos.

Neste projeto, pretendemos averiguar a hipótese acima discutida, de que o fato dos sistemas atuais não tratarem a tarefa de recuperação de informação como uma tarefa primariamente lingüística é a causa principal de não obterem melhorias significativas, pelo menos não em domínios abertos, com o uso de técnicas/recursos de PLN.

Este trabalho se propõe a desenvolver uma arquitetura lingüisticamente motivada para RI para português, mostrada na Figura 7. Por lingüisticamente motivada, entende-se que esta arquitetura será pensada sob o ponto de vista da língua portuguesa e de seu uso e que pretendemos assim utilizar dessas características, por nós empiricamente observadas, durante todo o processo de RI interpretação de consultas, criação de índices, mecanismo de comparação da consulta com o índice e apresentação dos resultados. Os usos citados acima podem ter sua motivação encontrada na Lingüística, no PLN ou na Ciência da Informação, ou até mesmo na Interação Homem-Computador, já que um sistema de RI envolve também um diálogo.

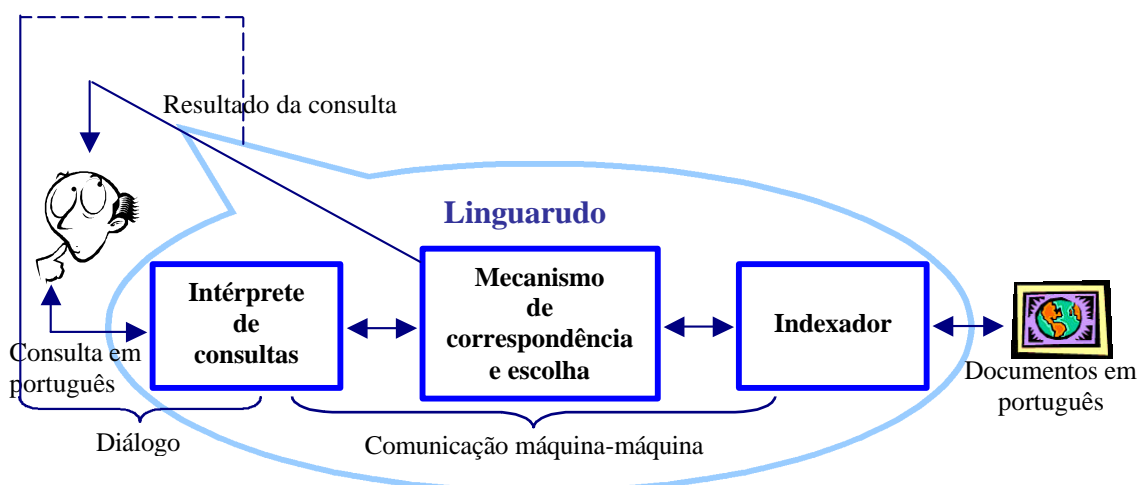


Figura 7 - Linguarudo _ Uma arquitetura lingüisticamente motivada para RI para português

5.1 Atividades Planejadas

A seguir delimitamos as atividades que serão desenvolvidas para a realização do trabalho aqui proposto. Essas atividades não seguem ordem cronológica. Algumas serão feitas em paralelo, conforme demonstra o cronograma na subseção seguinte.

1) Análise de textos e de consultas

Textos escritos têm muitas outras propriedades além de tratar sobre um ou mais tópicos, podem ser caracterizados de acordo com diferentes critérios, por exemplo, de acordo com o gênero (resumos, discussões, monólogos, persuasivos, notícias), de acordo com vários meios (podem ser ilustrados), com problemas lingüísticos (vagos, não gramaticais, ambíguos), com as necessidades do usuário em foco (difíceis, pequenos, repetitivos, focados, ofensivos, tendenciosos), e de acordo com o tempo (obsoletos). Muitas destas características são importantes para o usuário e algumas poderiam ser utilizadas na RI.

Assim como os textos, as consultas também poderiam ser caracterizadas de acordo com diferentes critérios originados dos diferentes tipos de necessidades de usuários, por exemplo: procura por respostas diretas para uma pergunta, procura a respeito de informações relacionadas a um determinado tópico, por notícias, por serviços on-line, por uma url de uma fonte de informação que já se conhece e por toda a informação disponível a respeito de um determinado assunto. O mesmo que foi dito a respeito dos textos é válido para as consultas. Não poderíamos escolher apenas uma estrutura como modelo de consulta padrão, a não ser que estivéssemos modelando um sistema de RI para um domínio e um tipo de necessidade de informação específicos.

Reconhecer as características de textos e consultas mais desejadas é um ponto essencial para modelarmos sistemas mais precisos e avaliarmos os sistemas de acordo com as reais necessidades do usuário. Faz parte desta atividade modelar e executar este estudo para definirmos inclusive os tipos de consultas que farão parte de nossa coleção de teste (Atividade 7).

2) Escolha do Modelo de Recuperação

Não é a proposta deste trabalho criar um novo Modelo de Recuperação, semelhante aos que foram tratados no Capítulo 2, Seção 2.1.3. Esta tarefa consiste em rever em mais detalhes as possibilidades de modelos de recuperação para definirmos o que mais se adequa à abordagem de RI que pretendemos criar. Dando principal atenção ao Modelo Lógico, proposto por van Rijsbergen (1986). Neste tipo de modelo se supõe que é possível representar o conteúdo de um documento por uma frase d e a necessidade de informação como foi apresentada na consulta por uma frase f . A verdade $d \models f$, significaria que a frase da consulta poderia ser inferida a partir da frase d , ou seja a informação capturada por d seria suficiente para inferir a informação representada por f . Huibers et al (1996) defendem o modelo lógico como um modelo que representaria bem a informação e seu fluxo, e mostram a teoria das situações (*situation theory*) (Huibers & Bruza, 1994) como sendo uma forma apropriada para modelar a RI.

3) Definição da arquitetura teórica linguisticamente motivada de RI

A arquitetura Linguarudo será composta de: intérprete de consultas (Quadro 1), mecanismo de correspondência e escolha (Quadro 2), indexador (Quadro 3) e diálogo que se dá através do intérprete de consultas, do mecanismo de correspondência e escolha e de uma interface cuidadosamente elaborada para possibilitar que o usuário possa elaborar sua consulta de forma mais eficiente.

Através do intérprete de consultas e da interface que será proposta pretende-se obter consultas não ambíguas, tanto determinando qual o assunto buscado por um usuário, quanto qual o enfoque que o usuário deseja que seja dado nas respostas. Enquanto que pela associação do intérprete de consultas, com o módulo de correspondência e escolha e com a interface pretendemos não só retornar os documentos que são relevantes e que seguem o estilo procurado pelo usuário, mas também: (i) possibilitar que o usuário saiba como compreender o funcionamento do sistema e assim se recuperar de erros do sistema; (ii) que o usuário possa ver que cometeu erros na elaboração de consultas; (ii) que o usuário possa mudar facilmente

de idéia quanto a seus objetivos durante as sessões de consultas e refiná-las ou modificá-las.

Quadro 1 – Módulo responsável pela interpretação das consultas

Intérprete de consultas	
Questão tratada	O que o usuário realmente deseja
Tarefas	<ol style="list-style-type: none"> 1. Investigação das possíveis interpretações para uma consulta 2. Para consultas ambíguas apresentar as opções para o usuário 3. Caso não consiga interpretar a consulta pedir ajuda do usuário
Possíveis técnicas	Análise morfofossintática Análise morfológica Análise sintática Corretor ortográfico Extração de múltiplos termos Extração de relações semânticas Tesouros e redes semânticas Uso de padrões de perguntas em língua portuguesa

Quadro 2 - Módulo responsável pela correspondência e escolha

Mecanismo de correspondência e escolha	
Questão tratada	Quais documentos estão relacionados à consulta, o quanto estão relacionados e como eles atendem a consulta
Tarefas	<ol style="list-style-type: none"> 1. Encontrar os documentos relevantes para um dado usuário 2. Mostrar os resultados para o usuário e auxiliá-lo a encontrar rapidamente a resposta que procura 3. Permitir que o usuário entenda qual a relação entre os documentos retornados
Possíveis técnicas	Características de estilo Classificação automática de textos quanto ao assunto Segmentação de textos Sumarização automática

Quadro 3 - Módulo responsável pela indexação

Indexador	
Questão tratada	A respeito do que trata um determinado documento
Tarefas	1. Identificar o idioma do documento 2. Identificar os assuntos tratados pelo documento
Possíveis técnicas	Análise morfológica Análise sintática Extração de múltiplos termos Extração de relações semânticas <i>Stemmer</i> Tesauros

Como pode ser visto pelos Quadros 1, 2 e 3, em nosso levantamento inicial de que recursos da língua portuguesa ou de quais ferramentas/recursos de PLN utilizar neste projeto não estão todos os recursos citados no Capítulo 3. A razão para isto, é que não estamos nos preocupando em utilizar recursos para minimizar o tamanho do índice, aumentar a velocidade de processamento das consultas ou aumentar o número de documentos recuperados. Ainda que algum destes fatos possa ocorrer, o que não seria indesejável, o objetivo principal desta arquitetura é poupar o usuário do contato com documentos que não são relevantes para sua consulta no momento em que foi feita. Por isso, estão nos quadros apenas os recursos/técnicas que levantamos a princípio como essenciais para cumprir este objetivo, ainda que alguns recursos possam facilitar também o aumento do número de documentos recuperados, como por exemplo, o uso de tesauros. É importante ressaltar, que é possível e provável que a visão inicial dos módulos mostrada nos quadros 1, 2 e 3 seja alterada durante este projeto. Em decorrência tanto da identificação de novas características que possam ser exploradas levantadas através da atividade 1, ou em decorrência da atividade 4.

Outra característica da arquitetura Linguarudo em destaque nos quadros anteriores é a importância do estudo de relacionamentos para este projeto. Como vimos no Capítulo 2, RI envolve identificar um subconjunto de documentos de uma coleção que provavelmente contenha informações relevantes em resposta a uma

consulta. Tipicamente, os sistemas de RI comparam as palavras-chave de um documento com os termos presentes nas consultas. Mas se as consultas contêm mais de um termo, então talvez também contenham relacionamentos semânticos entre os termos. Neste caso, os documentos relevantes para a consulta deveriam conter todos os termos das consultas e também os corretos relacionamentos entre eles. O fato de um sistema de RI passar a considerar e identificar corretamente os relacionamentos semânticos poderia melhorar sua precisão para algumas consultas eliminando os documentos que contêm os termos requeridos, mas não os relacionamentos desejados entre eles. Os tipos de relacionamento serviriam para identificar como lidar com conhecimento, se sabemos que dois termos estão relacionados como classe e instância, iremos tratá-los de forma diferente do que trataríamos se tivessem um relacionamento do tipo causa-efeito (Green et al, 2002).

Por estudar relacionamentos entenda estudar tipos de relacionamentos: hiponímia, troponímia, meronímia, causa-efeito (Green et al, 2002). E também estudar: (i) relacionamentos bibliográficos; relacionamentos entre textos e extra-textos, como por exemplo, citações e *hyperlinks*; (iii) relações entre tópicos e (iv) relacionamentos de relevância (Bean & Green, 2001).

4) Análise do uso de técnicas e ou recursos levantados como interessantes

É possível que algumas técnicas contribuam mais que outras para melhoria da precisão de um sistema de RI. É possível também que um dado recurso só seja útil, caso obtenha uma precisão mínima. Todas as técnicas e recursos que levantamos e levantarmos como interessantes terão de ser avaliados segundo estas duas perspectivas: (i) o recurso/técnica melhoraria consideravelmente a precisão de um sistema de RI; (ii) qual a precisão mínima que o recurso/ferramenta tem de alcançar para não causar mais problemas do que benefícios.

Da primeira análise podemos concluir, por exemplo, que muito menos é necessário para identificar o assunto de interesse de um usuário expresso por uma consulta, por exemplo, que apenas a análise morfológica, corretor ortográfico e uso de padrões já seriam suficientes. Da segunda análise poderíamos concluir, por exemplo, que caso o analisador sintático cometa mais de 10% de erros ele passa a ser

prejudicial para a precisão do sistema de RI. Ou poderíamos concluir, por exemplo, que o uso de *stemmer* aumenta a precisão do sistema de RI em 1% quando o *stemmer* é 100% preciso, e que por não termos disponível nem um *stemmer* mais de 70% preciso não deveríamos utilizar *stemmer*.

Ambas análises serão feitas manualmente, através de estudos empíricos. Cada técnica/recurso será analisado individualmente. No caso da primeira análise, iremos fazer a etiquetagem, classificação, extração ou substituição de termos manualmente para avaliarmos se as técnicas/recursos são úteis com 100% de acerto. Após a análise individual segue a análise das técnicas/recursos utilizados em conjunto.

A segunda análise partirá do valor da precisão das ferramentas/técnicas estado da arte para uma determinada tarefa. Por exemplo, verificar o que acontece dado que o melhor extrator de relações semânticas para português disponível atinja apenas 70% de precisão introduzindo 30% de erro nos testes de seu uso na RI (feitos na primeira tarefa manualmente).

Com os resultados das duas análises saberemos avaliar o custo benefício da inclusão de cada técnica ou recurso na arquitetura e assim decidir por sua inclusão ou não na mesma. Ou seja, esta atividade é essencial para atividade 3 e por isso ambas acontecem concomitantemente. Por isso, é necessário que antes mesmo de termos um protótipo de teste (atividade 8), tenhamos uma coleção de textos/consultas de teste.

5) Definição das características do protótipo de teste

Com o objetivo de testarmos a abordagem proposta, iremos desenvolver o protótipo de uma ferramenta de busca. A princípio pensamos no protótipo de uma máquina de busca, mas a idéia foi rejeitada posteriormente por acreditarmos que, mesmo que fique comprovado um desempenho superior adotando-se a abordagem proposta, não acreditamos que a tarefa de busca fosse ser executada de forma tão rápida como é requisito para este tipo de aplicação. Optamos então por uma aplicação que funcionasse de certa forma como uma ferramenta de meta-busca, compondo sua base de documentos no momento da busca através dos resultados da consulta aplicada

(enviada) a diferentes máquinas de busca. Esta atividade consiste da modelagem deste protótipo que depende em parte da realização da Atividade 6.

6) Avaliação de máquinas de busca para português

Com esta atividade pretendemos avaliar como as máquinas de busca estão atendendo as necessidades dos usuários que procuram por informações em português, utilizando-nos da estratégia definida na Atividade 7.

7) Definição da estratégia de avaliação e coleção de testes

A estratégia de busca e coleção de teste considerará as necessidades atuais dos usuários, como o encontro de notícias e serviços on-line. Já que este tipo de avaliação é o que realmente importa atualmente, sendo reforçado por Hawking, no Tutorial de avaliação de máquinas de busca no SIGIR 2002. A coleção de teste será utilizada nas atividades 4, 6 e 8.

Outra característica importante da coleção de teste é a presença tanto de consultas em linguagem natural, como de consultas através de palavras-chave. Assim poderemos testar todas as funcionalidades da arquitetura, tanto a interação através da entrada de consultas em linguagem natural (tipo 1 que será permitido), quanto através de palavras e não frases em língua natural, já que temos que considerar o fato de que por mais que a entrada em língua natural aumente a precisão, existirão aplicações para as quais este tipo de entrada não são práticos, como por exemplo, os sistemas de RI para web utilizados em telefones celulares. Esperamos, se possível, contar nesta atividade não só com os resultados da Atividade 1, mas também com a ajuda de um profissional de Ciência da Informação.

8) Desenvolvimento e avaliação do protótipo de ferramenta de busca

9) Escrita e defesa da Tese

10) Publicações

A elaboração de relatórios técnicos e artigos para conferências nacionais e internacionais é também objetivo deste trabalho e deve ocorrer conforme seu progresso. Por isso esta atividade não consta no cronograma da próxima subseção.

5.2 Cronograma

Meses	Atividades								
	I	II	III	IV	V	VI	VII	VIII	IX
Abr2003	X								
Mai2003	X								
Jun2003	X						X		
Jul2003							X		
Ago2003		X					X		
Set2003		X	X	X		X			
Out2003		X	X	X	X				
Nov2003			X	X					
Dez2003			X	X					
Jan2004			X	X					
Fev2004			X	X					
Mar2004			X	X					
Abr2004			X	X					
Mai2004								X	
Jun2004								X	
Jul2004								X	
Ago2004								X	
Set2004								X	
Out2004								X	
Nov2004									X
Dez2004									X
Jan2005									X
Fev2005									X
Mar2005									X

5.3 Metodologia

Como dito anteriormente, acreditamos que um dos problemas do uso de PLN na RI não ter tido êxito é o fato de tradicionalmente o PLN ter sido aplicado a sistemas e arquiteturas já prontas e que não foram pensados desde o princípio tendo em vista a exploração de características da língua. Por isso adotamos neste trabalho uma estratégia diferenciada. Não iremos propor modificações a uma arquitetura já existente, implementar tais modificações e então testar se as modificações resultaram em melhorias, mas sim propor uma arquitetura considerando desde o início a língua portuguesa. Como pode ser visto através das atividades listadas na seção 5.1 a metodologia deste projeto é uma metodologia exploratória. Pois tendo em mente as possíveis técnicas/recursos/ferramentas que podemos utilizar iremos avaliá-las em textos e consultas que compõem uma coleção de teste e, a partir dos resultados desta avaliação, propor a arquitetura Linguarudo.

5.4 Contribuições

Temos consciência que por limitações de tempo talvez não possamos definir formalmente o uso de todas as técnicas e recursos lingüísticos que seriam úteis. No entanto, esperamos contribuir com uma abordagem para RI para português que por mais limitada que seja por questão de tempo, terá sido pensada, ainda que de forma mais simplificada, em todas as fases do processo de RI segundo o uso da língua portuguesa.

5.5 Viabilidade e Recursos Disponíveis

Este projeto será realizado parte no SINTEF, dentro do projeto Linguateca e, parte no NILC. Ambos dispõem de ótimas condições de laboratório computacional e recursos/ferramentas de PLN e material para informação sobre estudos lingüísticos do português. O projeto Linguateca se preocupa com a reunião de recursos para o português desde 1998 (Santos, 2002). E o NILC possui experiência no desenvolvimento de sistemas e recursos de PLN para português desde 1993 e conta também com uma equipe interdisciplinar formada por cientistas da computação

e lingüistas. Os projetos Diadorim¹³, GOSPELL¹⁴, Explosa¹⁵, Lacio Web¹⁶, Wordnet.Br¹⁷, Dizer¹⁸, Supor¹⁹, TEP²⁰ e Curupira²¹ são exemplos de projetos em andamento ou concluídos no NILC que podem ter as ferramentas/recursos produzidos neles utilizados ou servindo de inspiração para partes da arquitetura que será proposta neste projeto.

¹³ <http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm>

¹⁴ <http://www.nilc.icmc.usp.br/nilc/projects/gospell.htm>

¹⁵ <http://www.dc.ufscar.br/~lucia/PROJECTS/EXPLOSA.htm>

¹⁶ <http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm>

¹⁷ <http://www.nilc.icmc.usp.br/nilc/projects/wordnetbr.htm>

¹⁸ <http://www.nilc.icmc.usp.br/~thiago/DiZer.html>

¹⁹ <http://www.dc.ufscar.br/~mmodolo/Project.htm>

²⁰ <http://www.nilc.icmc.usp.br/nilc/tools/tep.htm>

²¹ <http://www.nilc.icmc.usp.br/nilc/tools/curupira.html>

Bibliografia e referências

(Abdulla et al, 1997) Abdulla, G.; Liu, B.; Saad, R.; Fox, E. 1997. Characterizing WWW queries. Computer Science Department, Virginia Tech, Technical Report, TR-97-04. Disponível em http://historical.ncstrl.org/tr/ps/vatech_cs/TR-97-04.ps

(Aires & Santos, 2002) Aires, Rachel Virgínia Xavier; Santos, Diana. 2002. Measuring the Web in Portuguese. Euroweb 2002. Oxford, Inglaterra.

(Aires, 2000) Aires, Rachel Virgínia Xavier Aires. 2000. Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil. Dissertação de mestrado, Instituto de Ciências Matemáticas de São Carlos – USP. Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/mestradorachel.html>.

(Allan, 2001) Allan, Keith. Natural Language Semantics. Blackwell Publishers. 2001.

(Arampatzis et al, 1999) Arampatzis, A. T.; van der Weide, Th. P.; van Bommel, P.; Koster, C. H. A. 1999. Linguistically Motivated Information Retrieval. Disponível em <http://citeseer.nj.nec.com/arampatzis00linguistically.html>.

(Ardizzone & La Casia, 1997) Ardizzone, E.; La Casia, M. 1997. Automatic Video Database Indexing and Retrieval. Multimedia Tools and Applications, Vol. 4, No. 1, p. 29-56.

(Baeza-Yates & Ribeiro-Neto, 1999) Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley. 1999.

(Bean & Green, 2001) Bean, Carol A.; Green, Rebecca. Relationships in the Organization of Knowledge. Kluwer Academic Publishers. 2001.

(Becker & Hayes, 1963) Becker, Joseph; Hayes, Robert Mayo. Information storage and retrieval: tools, elements, theories. New York, Wiley. 1963.

(Belew, 2000) Belew, Richard K. Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press. 2000.

(Bick, 2000) Bick, Eckhard. The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press. 2000.

(Bowles, 1998) Bowles, Mark D. The Information Wars: Two Cultures and the Conflict in Information Retrieval, 1945–1999. Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems. Disponível em: http://www.chemheritage.org/HistoricalServices/ASIS_documents/ASIS98_Bowles.pdf

(Bräscher, 1999) BRÄSCHER, M. 1999. Tratamento automático de ambigüidades na recuperação da informação. Tese de Doutorado em Ciência da Informação –

Universidade de Brasília.

(Bräscher, 2002) Bräscher, Marisa. A ambigüidade na Recuperação de Informação. DataGramaZero - Revista da Ciência da Informação - v.3 n.1 fev 2002. Disponível em: http://www.dgzero.org/Atual/Ind_onum.htm

(Brown & Yule, 1983) BROWN, Gillian; Yule, George. Discourse analysis. Cambridge University Press, 1983.

(Bush, 1945) Bush, Vannevar. As We May Think. The Atlantic Monthly; Julho, 1945. Volume 176, No. 1; p. 101-108. Disponível em <http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml>.

(Cacheda & Viña, 2001a) Cacheda, F; Viña, Á. 2001. Understanding how people use search engines: a statistical analysis for e-Business. In: Proceedings of the e-2001 (e-Business and e-Work Conference and Exhibition), 1, p. 319-325. Disponível em <http://citeseer.nj.nec.com/496769.html>.

(Cacheda & Viña, 2001b) Cacheda, F.; Viña, Á. 2001. Experiences retrieving information in the world wide web. In: Proceedings of the 6th IEEE Symposium on Computers and Communications, p. 72-79. Disponível em <http://citeseer.nj.nec.com/488520.html>.

(Chandrasekar & Srinivas, 1997) Chandrasekar, R.; Srinivas, B. Gleaning information from the Web: Using Syntax to Filter out Irrelevant Information. In Proceedings of 1997 AAAI Spring Symposium. Technical Report SS-97-02.

(Chen, 1994) Chen, Hsinchun. 1994. Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. Journal of the American Society for Information Science, 46(3), p. 194-216. Disponível em <http://ai.bpa.arizona.edu/papers/>.

(Chiaramella et al, 1996) Chiaramella, Yves; Mulhem, Philippe; Fourel, Franck. A Model for Multimedia Information Retrieval. 1996. Technical Report Fermi ESPRIT BRA 8134, University of Glasgow. Disponível em <http://citeseer.nj.nec.com/cache/papers/cs/1764/http://zSzzSzoutlet.imag.fr/zSzfourelzSzpubzSzfermi96zSzReport-MRIM.pdf/chiaramella96model.pdf>

(Chu & Rosenthal, 1996). Heting Chu; Marilyn Rosenthal. Search Engines the World Wide Web: A comparative study and evaluation methodology. ASIS 1996. <http://www.asis.org/annual-96/ElectronicProceedings/chu.html>

(Cleverdon, 1962) Cleverdon, Cyril W. 1962. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield Coll. of Aeronautics, Cranfield, Inglaterra

(Cole, 1998) Cole, Charles. 1998. Intelligent Information Retrieval: Diagnosing Information Need. Part 1. The Theoretical Framework for Developing an Intelligent IR Tool. Information Processing & Management. Vol.34. N°6. p. 709-720.

(Cooper, 1968) Cooper, W. S. 1968. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. Journal of

the American Society for Information Science, 19, 30-41.

(Crivellari & Melucci, 2000) Crivellari, Franco; Melucci, Massimo. 2000. Web Document Retrieval using Passage Retrieval, Connectivity Information, and Automatic Link Weighting — TREC-9 Report. Disponível em http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

(Cunha, 1997) Cunha, Cecília Kremer Vieira da. 1997. Planejador de Respostas Explicativas Baseado em uma Biblioteca de Esquemas RST. Dissertação de Mestrado. Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro. Disponível em http://peirce.inf.puc-rio.br/serg/pub/ceciliak/DISSERT_PDF.pdf

(Deerwester et al 1990) Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K.; Harshman, Richard. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41:391-407.

(Dias & Nunes, 2001) Dias, Gaël; Nunes, Segio. 2001. Combining Evolutionary Computing and Similarity Measures to Extract Collocations from Unrestricted Texts. Proceedings of RANLP - 2001 (Recent Advances in NLP), p. 5-7.

(Dias et al, 1999) Dias, Gaël; Lopes, José Gabriel Pereira; Guilleré, Sylvie. 1999. Mutual Expectation: A Measure for Multiword Lexical Unit Extraction. Vextal 99: Venezia per il Trattamento Automatico delle Lingue.

(Estopà Bagot, 2001) Estopà Bagot, R. 2001. Extracción de Terminologia: elementos para la construcción de un extractor. In TradTerm 7. Revista do Centro Interdepartamental de Tradução e Terminologia FFLCH - USP, p. 225-50.

(Faloutsos & Oard, 1995) Faloutsos, Christos; Oard, Douglas. 1995. A Survey of Information Retrieval and Filtering Methods. Technical Report, University of Maryland at College Park. Disponível em <http://www.citeseer.nj.nec.com/faloutsos96survey.html>.

(Fayyad, 1997) Fayyad, U. 1997. Editorial. Data Mining and knowledge Discovery. 1:5-10.

(Feldman, 1999) Feldman, Susan. 1999. NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. Online, Inc. Disponível em <http://www.onlineinc.com/onlinemag/OL1999/feldman5.html>.

(Frakes & Baeza-Yates, 1992) Frakes, W. B.; Baeza-Yates, R. Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1992.

(Gamallo et al, 2002) Gamallo, Pablo; Gonzalez, Mario; Agustini, Alexandre; Lopes, Gabriel; Lima, Vera Lucia Strube de. 2002. Mapping Syntactic Dependencies onto Semantic Relations. In ECAI'02, Workshop on Natural Language Processing and Machine Learning for Ontology Engineering, p. 15-22. Disponível em <http://www.inf.pucrs.br/~gonzalez/docs/art-ecai.pdf>

(Gasperin, 2001) Gasperin, Caroline Varaschin. 2001. Extração automática de relações semânticas a partir de relações sintáticas. Dissertação de Mestrado.

Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

(Gonzalez & Lima, 2001a) Gonzalez, Marco; Lima, Vera Lúcia Strube de. Recuperação de Informação e Expansão Automática de Consulta com Theusarus: uma avaliação. 2001. XXVII Conferencia Latinoamericana de Informatica (CLEI'2001).

(Gonzalez & Lima, 2001b) Gonzalez, Marco; Lima, Vera Lúcia Strube de. Semantic Thesaurus for Automatic Expanded Query in Information Retrieval. 2001. IEEE Computer Society Press. 8th International Symposium on String Processing and Information Retrieval (SPIRE'2001), p.68-75.

(Gonzalez & Lima, 2001c) Gonzalez, Marco; Lima, Vera Lúcia Strube de. T-Lex: Thesaurus com Estruturação Semântica e Operações Gerativas. 2001. Thesaurus com Estruturação Semântica e Operações Gerativas. XXVII Conferencia Latinoamericana de Informatica (CLEI'2001).

(Gordon & Pathak, 1999) GORDON, Michael; Pathak, Praveen. Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing and Management 35, 1999, 141-180.

(Green et al, 2002) Green, Rebecca; Bean, Carol A.; Myaeng, Sung Hyon. The Semantics of Relationships: An Interdisciplinary Perspective. Kluwer Academic Publishers. 2002.

(Grefenstette & Nioche, 2000) Grefenstette, Gregory; Nioche, Julien. 2000. Estimation of English and non-English Language Use on the WWW. In: RIAO'2000. Disponível em <http://www.xrce.xerox.com/competencies/content-analysis/publications/Documents/P19137/content/RIAO2000gref.pdf>

(Greisdorf & Spink, 2001) Greisdorf, Howard; Spink, Amanda. Median Measure: an approach to IR systems evaluation. Information Processing and Management 37, 843-857. 2001.

(Gwizdka & Chignell, 1999) Gwizdka, Jacek; Chignell, Mark. Towards Information Retrieval Measures for Evaluation of Web Search Engines. Unpublished manuscript. http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf

(Hawking et al, 1999) Hawking, David; Craswell, Nick; Harman, Donna. 1999. Results and Challenges in Web Search Evaluation. WWW8. <http://www8.org/w8-papers/2c-search-discover/results/results.html>

(Hawking et al, 2000) Hawking, David; Craswell, Nick; Bailey, Peter; Griffiths, Kathy. Measuring Search Engine Quality. Journal of Information Retrieval. <http://www.wkap.nl/journalhome.htm/1386-4564>.

(Hawking, 2002) Hawking, David. Web Search Evaluation. Tutorial on Search from the Web to the Enterprise: Issues, Solutions, Evaluation. SIGIR 2002.

(Hearst, 1997) Hearst, Marti A. 1997. Text Data Mining – Issues, Techniques, and the Relation to Information Access. UW/MS Workshop on Data Mining. Disponível

em <http://www.sims.berkeley.edu/~hearst/talks/dm-talk/>.

(Hiemstra, 2001) Hiemstra, Djoerd. Using Language Models for Information Retrieval. 2001. Tese de doutorado, Centre for Telematics and Information Technology, University of Twente. Disponível em <http://wwwhome.cs.utwente.nl/~hiemstra/papers/>

(Huibers & Bruza, 1994) Huibers, T. W. C.; Bruza, P. D. 1994. Situations, a General Framework for Studying Information Retrieval. Information retrieval: In: Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialists Group. Disponível em <http://citeseer.nj.nec.com/huibers94situations.html>.

(Huibers et al, 1996) Huibers, T. W. C.; Lalmas, M.; van Rijsbergen, C. J. Information 1996. Retrieval and Situation Theory. In: Proceedings of SIGIR 1996. Disponível em <http://portal.acm.org/citation.cfm?id=381986&coll=GUIDE&dl=GUIDE&ret=1#Fulltext>.

(Inc. Magazine, 1999) Inc. Magazine, Janeiro de 1999. Data Data. <http://www.inc.com/magazine/19990101/715.html>.

(Jansen & Pooch, 2000) Jansen, B. J., Pooch, U. 2000. A review of web searching studies and a framework for future research. In: Journal of the American Society of Information Science and Technology, 523, p. 235 – 246. Disponível em <http://citeseer.nj.nec.com/417587.html>.

(Jansen & Spink, 2000) Jansen, B. J., Spink, A. 2000. The Excite Research Project: A study of searching characteristics by web users. In: ASIS Bulletin, 27 1. Disponível em <http://citeseer.nj.nec.com/415792.html>.

(Jansen et al, 1998) Jansen, B.; Spink, A.; Bateman, J.; Saracevic, T. 1998. Real Life Information Retrieval: A Study of User Queries on The Web. In: SIGIR 98, 321, 5-17. Disponível em <http://jimjansen.tripod.com/academic/acad.html#ResP>.

(Jansen et al, 2000) Jansen, B; Spink, A.; Saracevic, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web, Information Processing and Management 362, 207-227. Disponível em <http://jimjansen.tripod.com/academic/acad.html#ResP>

(Jansen, 2000) Jansen, B. J. 2000. An investigation into the use of simple queries on web IR systems. Information Research: an international electronic journal, 61. Disponível em <http://citeseer.nj.nec.com/420204.html>

(Jardine & van Rijsbergen, 1971) Jardine, N.; Van Rijsbergen, C. J. 1971. "The use of hierarchic clustering in information retrieval". Information Storage and Retrieval, 7, p. 217-240.

(Jing & Croft, 1994) Jing, Yufeng; Croft, W. Bruce. An Association Thesaurus for Information Retrieval. 1994. Proceedings of RIAO-94, 4th International Conference ``Recherche d'Information Assistee par Ordinateur''. Disponível em <http://www>.

www.cs.umass.edu/Dienst/UI/2.0/Describe/ncstrl.umassa_cs/

(Sparck Jones, 1964) Sparck Jones, Karen. Synonymy and Semantic Classification. Tese, Cambridge, 1964.

(Sparck Jones, 1999) Sparck Jones, Karen. 1999. What is the role of NLP in Text Retrieval?. In Tomek Strzalkowski, editor, Natural Language Information Retrieval. Kluwer, Boston.

(Karlgrén, 1999) Karlgrén, Jussi. Non-topical factors in information access. Invited talk to webnet'99. Disponível em <http://www.sics.se/~jussi/papers>

(Kaszkiel & Zobel, 1997) Kaszkiel, M.; Zobel, J. 1997. Passage retrieval revisited. In SIGIR 1997, p. 178 -185.

(Katz, 1996) Katz, Slava. Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2:15-60.

(Kirsch, 1998) Kirsch, S. 1998. "Infoseek's experiences searching the internet", In: SIGIR 98.

(Koch, 1998) Koch, Ingedore Villaça. A coesão textual. Coleção Repensando a língua portuguesa. Editora Contexto. 10ª edição. São Paulo, 1998.

(Koenemann & Belkin, 1996) Koenemann, Jürgen; Belkin, Nicholas J. 1996. A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. SIGIR 1996. p. 205-212.

(Kraaij & Westerveld, 2000) Kraaij, Wessel; Westerveld, Thijs. TREC-9: How different are Web documents? 2000. Disponível em http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

(Kuramoto, 2002) Kuramoto, Hélio. Sintagmas Nominais: uma nova propostas para a recuperação de informação. DataGramaZero - Revista da Ciência da Informação - v.3 n.1 fev 2002. Disponível em: http://www.dgzero.org/Atual/Ind_onum.htm

(Lassila & McGuinness, 2001) Lassila, Ora; McGuinness, Deborah. The Role of Frame-Based Representation on the Semantic Web. Linköping Electronic Articles in Computer and Information Science. Vol.6. 2001. Disponível em <http://www.ida.liu.se/ext/epa/cis/2001/005/tcover.html>.

(Lawrence & Giles, 1998) Lawrence, S.; Giles, C. Searching the Word Wide Web. Science, 280 Apr.3, 1998, 98-100.

(Lawrence & Giles, 1999) Lawrence, S., Giles, C. L. 1999. Accessibility of information on the Web. Nature, 400 July 8, 1999, p. 107-109. Disponível em: <http://www.neci.nec.com/~lawrence/papers.html>

(Lawrence, 2000) Lawrence, Steve. Context in Web Search. IEEE Data Engineering Bulletin, Vol. 23, Nº3, 25-32, 2000. Disponível em <http://citeseer.nj.nec.com/lawrence00context.html>

(Lee, 1995) Lee, J. H. (1995). Analyzing the effectiveness of extended boolean models in information retrieval. Technical Report TR95-1501, Cornell University. Disponível em <http://cs-tr.cs.cornell.edu/>

(Lesk, 1995) Lesk, Michael. The Seven Ages of Information Retrieval. In As We May Think: A 50th Anniversary Celebration of Bush's Vision, MIT, Outubro 1995. Disponível em <http://www.ifla.org/VI/5/op/udtop5/udtop5.htm>.

(Lewis & Sparck Jones, 1996) Lewis, David D.; Sparck Jones, Karen. Natural Language Processing for Information Retrieval. Communications of the ACM, Vol. 39, Nº1, 92-101, Janeiro de 1996. Disponível em <http://citeseer.nj.nec.com/86648.html>.

(Lewis, 1996) Lewis, David. Dying for Information: An Investigation Into the Effects of Information Overload in the USA and Worldwide. London: Reuters Limited, 1996.

(Lopes & Quaresma, 1999) Lopes, José Gabriel P.; Quaresma, Paulo. 1999. A Dialog System for controlling question/answer dialogues. In Potapova, Text Processing and Cognitive Technologies, vol. 2, Moscovo, Rússia, p.75-86.

(Lopes & Rodrigues, 1996) Lopes, José Gabriel P.; Rodrigues, Irene Pimenta. 1996. Abductive Reasoning Applied to Text Processing: Retrieval of Temporal Information. In Dahl, Veronica & A. Sobrino (eds.), Estudios sobre Programación Lógica y sus aplicaciones. Publicacións da Universidade de Santiago de Compostela.

(Losada & Barreiro) Losada, David E; Barreiro, Álvaro. Retrieval Situations and Belief Change. 11th International Workshop on Database and Expert Systems Applications. 2000. Disponível em <http://www.computer.org/proceedings/dexa/0680/06800531abs.htm>.

(Loudon, 2002) Loudon, Gareth; Sacher, Heiko; Kew, Leong Mun. Design Issues for Mobile Information Retrieval. Workshop: Mobile Personal Information Retrieval. SIGIR 2002. 2002.

(Luhn, 1958) Luhn, H.P., 'The automatic creation of literature abstracts', IBM Journal of Research and Development, 2, 159-165 (1958).

(Luhn, 1959) Luhn, H.P. 1959. "Auto-encoding of documents for information retrieval." Disponível em: <http://web.utk.edu/~jgant/hanspeterluhn.html>

(Lyman et al, 2000) Peter Lyman, Hal R. Varian, James Dunn, Aleksey Strygin, Kirsten Swearingen. 2000. How Much Information? Extraído de <http://www.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf>

(Maron & Kuhns, 1960) Maron, Melvin Earl; Kuhns, J. L. On relevance, probabilistic indexing, and information retrieval. Journal of the ACM, 7(3):216-244,

Julho 1960.

(Martin, 1995) Martin, Joel D. Clustering Full Text Documents. In IJCAI-95 Workshop on Data Engineering for Inductive Learning. 1995. Disponível em <http://citeseer.nj.nec.com/martin95clustering.html>

(Martins et al, 2002) Martins, R. T.; Hasegawa, R.; Nunes, M.G.V. Curupira: um parser funcional para o português. NILC-TR-02-26, Dezembro 2002. Disponível em: <http://www.nilc.icmc.usp.br/nilc/publications.htm#JournalArticles>.

(Miorelli, 2001) Miorelli, Sandra Terezinha. 2001. ED-CER - Um Método para Extração do Sintagma Nominal em Sentenças em Português. Dissertação de Mestrado. Instituto de Informática da Pontifícia Universidade Católica do Rio Grande do Sul. Disponível em <http://www.pucrs.br/inf/pos/dissertacoes/arquivos/sandra.pdf>

(Notess, 1999) Greg R. Notess. Comparing Internet Search Engines. <http://www.csu.edu.au/special/online99/proceedings99/103a.htm>

(Notess, 2000) Greg R. Notess. Search Engine Statistics: Dead links report. <http://www.notess.com/search/stats/deads.shtml>.

(Notess, 2002) Greg R. Notess. Search Engine Statistics: Unique Hits Report. <http://www.notess.com/search/stats/unique.html>.

(Oard, 1997) Oard, Douglas W. Cross-Language Information Retrieval Bibliography. 1997. Disponível em: <http://citeseer.nj.nec.com/oard97crosslanguage.html>.

(Ohlman, 1998) Ohlman, Herbert. Mechanical Indexing: A Personal Remembrance. Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems. Disponível em: http://www.chemheritage.org/HistoricalServices/ASIS_documents/ASIS98_Ohlman.pdf

(Padilha, 1997) Padilha, Emiliano Gomes. 1997. Interpretação Temporal: Representação e Raciocínio. Dissertação de Mestrado. Instituto de Informática da Universidade Federal do Rio Grande do Sul (UFRGS). Disponível em: <http://www.iccs.informatics.ed.ac.uk/~emilianp/works/Dissmest.zip>

(Page et al, 1998) Page, Larry; Brin, Sergey; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. (1998). Stanford Digital Library Technologies Project. Disponível em: <http://citeseer.nj.nec.com/page98pagerank.html>

(Paice, 1984) Paice, C. P. (1984). Soft evaluation of boolean search queries in information retrieval systems. Information Technology: Research and Development 3 (1), 33-42.

(Paraboni, 1997) Paraboni, Ivandré. 1997. Uma arquitetura para a Resolução de Referências Pronominais Possessivas no Processamento de Textos em Língua Portuguesa. Dissertação de Mestrado. Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul. Disponível em:

<http://www.inf.pucrs.br/ppgcc/dissertacoes/arquivos/ivandre.zip>

(Pardo et al, 2002) Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. 2002. Extractive summarization: how to identify the gist of a text. In the Proceedings of the International Information Technology Symposium - I2TS. Florianópolis-SC, Brazil. Also published in M.G.V. Nunes and A.C.P.L.F. Carvalho (eds.), *Cadernos de Computação*, Vol. 3, N. 2. Disponível em: <http://www.nilc.icmc.usp.br/~thiago/>

(Pelizzoni, 2002) Pelizzoni, J.M. *Preâmbulo ao aconselhamento ortográfico para o português do Brasil - Uma releitura baseada em utilidade e conhecimento lingüístico*. Tese de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. Apr, 2002

(Perkins, 2003) Perkins, Alan. *The Classification of Search Engines Spam*. Disponível em <http://www.ebrandmanagement.com/whitepapers/spam-classification/>. 2003.

(Peters, 2000) Peters, C. (Ed.), 2000. *Cross-language Information Retrieval – Revised papers of the Workshop of the Cross-language Information Retrieval Forum, CLEF 2000*, Lisboa, Portugal, In: LNCS 2069.

(Plank, 2002) Plank, Terry. *How Search Engines Look at Links*. Search Day, 13 de Junho de 2002. Disponível em <http://searchenginewatch.com/searchday/02/sd0613-links.html>.

(Ponte & Croft, 1998) Ponte, J. M.; Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*.

(Pratt & Fagan, 2000) Pratt, Wanda; Fagan, Lawrence. The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association*, Vol 7, 6, 2000. Disponível em <http://www1.ics.uci.edu/~pratt/main.html>

(Ribeiro & Muntz, 1996) Ribeiro, B. A. N.; Muntz, R. 1996. A belief network model for ir. In *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*, p. 252–260.

(Ribeiro et al, 1998) Ribeiro, Ana Paula; Fonseca, Rodrigo; Meira Jr., Wagner; Almeida, Virgílio. 1998. *Classificação Semântica Automática de Documentos da WWW*. In *Actas Eletrônica do I Workshop sobre Fatores Humanos em Sistemas Computacionais: Compreendendo Usuários, Construindo Interfaces*. p. 131-137.

(Robertson & Sparck Jones, 1996) Robertson, S. E.; Sparck Jones, Karen. 1996. Simple, proven approaches to text-retrieval. Technical Report 356, Computer Laboratory. University of Cambridge.

(Robertson & Teather, 1974) S.E. Robertson; D. Teather. A statistical analysis of retrieval tests: a Bayesian approach. *Journal of Documentation*, 30, 273-282. 1974.

(Rocha, 1999) Rocha, Marco. 1999. A corpus-based study of anaphora in English and Portuguese. In Botley, Simon & McEnery, A.M.(eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, *Studies in Corpus Linguistics* 3.

p.81-94. Amsterdam: John Benjamins Publishing Company.

(Salton & McGill, 1983) Salton, G.; McGill, M. J. (Eds.) (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

(Salton et al, 1975a) Salton, G; Yang, C.S.; Yu, C.T. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 36:33--44, 1975.

(Salton et al, 1975b) Salton, G.; Wong, A; Yang, C. S. A vector space model for automatic indexing. Communications of the ACM, 18:613-620, 1975.

(Salton et al, 1983) Salton, G.; Fox, E. A.; Wu, H. 1983. Extended boolean information retrieval. Communications of the ACM 26 (11), p. 1022--1036.

(Salton, 1968) Salton, Gerard. Automatic Information Organization and Retrieval. McGraw-Hill. 1968.

(Salton, 1975) Salton, G. 1975. A Theory of Indexing, The Society for Industrial and Applied Mathematics.

(Sanderson & Croft, 1999) Sanderson, Mark; Croft, Bruce. Deriving concept hierarchies from text. SIGIR 1999. <http://portal.acm.org/citation.cfm?id=312679&coll=portal&dl=ACM&ret=1#Fulltext>

(Sanderson, 1994) Sanderson, Mark. Word Sense Disambiguation and Information Retrieval. SIGIR 1994. Disponível em http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/SIGIR94.pdf.

(Sant'Anna, 2000) Sant'Anna, Victor Martins. 2000. Cálculo de Referências Pronominais Demonstrativas na Língua Portuguesa Escrita. Dissertação de Mestrado. Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul. Disponível em <http://www.inf.pucrs.br/ppgcc/dissertacoes/arquivos/victor.zip>

(Santos, 2002) Santos, Diana. "Um Centro de Recursos para o Processamento Computacional do português", DataGramZero, v.3 n.1 fev/02. Disponível em http://www.dgz.org.br/fev02/Art_02.htm

(Saracevic, 1995) T. Saracevic. 1995. Evaluation of evaluation in information retrieval. Proceedings of SIGIR 95, 138-146. Disponível em <http://www.scils.rutgers.edu/~muresan/Docs/sigirSaracevic1995.pdf>

(Schatz et al, 1996) Schatz, B. R.; Johnson, E. H.; Cochrane, P.A. Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval. Proceedings of Digital Libraries '96 (Bethesda MD, March 1996). ACM Press.126-133.

(Schatz, 1997). Schatz, Bruce R. Information Retrieval in Digital Libraries: Bringing Search to the Net. Science, V.275, 1997, p. 327-334. Disponível em <http://citeseer.nj.nec.com/schatz97information.html>.

(Shehory, 1999) Shehory, Onn. Spawning information agents on the web, Intelligent Information Agents, M. Klusch (Ed.), Springer 1999. Disponível em <http://www.citeseer.nj.nec.com/198201.html>

(Silva, 2001) Silva, Gilberto F. 2001. Representação do Léxico para Reconhecimento da Similaridade de Palavras no Português. Dissertação de Mestrado. Departamento de Engenharia de Sistemas do Instituto Militar de Engenharia. Maio de 2001. Disponível em <http://ipanema.ime.eb.br/~de9/teses/2001/Gilberto.zip>

(Smadja, 1993) Smadja, Frank. Retrieving Collocations from Text: XTRACT. Computational Linguistics, 19:143-177.

(Smeaton, 1990) Smeaton, A. F. Introduction: Natural Language Processing and information retrieval. Information Processing and Management, v. 26, n.1 p. 19-20, 1990.

(Smeaton, 1991) Smeaton, A. F. Prospects for intelligent, language-based information retrieval. Online Reviewm, v. 15, n.6, p. 373-382, 1991.

(Spink et al, 2002) Spink, A.; Jansen, B. J.; Wolfram, D.; Saracevic, T. 2002. From E-sex to E-commerce: Web Search Changes. IEEE Computer. 35(3), 107 - 111. Disponível em http://jimjansen.tripod.com/academic/pubs/ieee_computer.pdf.

(Storb & Wazlawick, 1998) Storb, B. H.; Wazlawick, R. S. Um modelo de Recuperação de Documentos para a Língua Portuguesa utilizando Stemming Difuso. PROPOR 1998.

(Strzalkowski et al, 1994) Strzalkowski, Tomek; Carballo, Jose Perez; Marinescu, Mihnea. Natural Language Information Retrieval: TREC-3 REPORT. 1994. Disponível em <http://citeseer.nj.nec.com/51110.html>.

(Strzalkowski et al, 1997) Strzalkowski, Tomek; Lin, Fang; Perez-Carballo, Jose. Natural Language Information Retrieval: TREC-6 REPORT. 1997. Disponível em <http://citeseer.nj.nec.com/90739.html>.

(Strzalkowski et al, 1999) Strzalkowski, Tomek; Perez-Carballo, Jose; Karlgren, Jussi; Hulth, Anette; Tapanainen, Pasi; Lahtinen, Timo. Natural Language Information Retrieval: TREC-8 REPORT. 1999. Disponível em <http://trec.nist.gov/pubs/trec8/papers/index.track.html>.

(Su et al, 1998) L. T. Su; H. Chen; X. Dong. Evaluation of Web-based search engines from the end-user's perspective: a pilot study. Proceedings of the Annual Conference for the American Society for Information Science, 348-361.

(Su, 1998) SU, Louise T. Value of search results as a whole as the best measure of information retrieval performance. Information Processing and Management Vol.34, nº 5, 557-579. 1998.

(Swets, 1963) J. A. Swets. Information Retrieval Systems. Science, 141, 245-250. 1963.

(Tombros & Sanderson) Tombros, Anastasios; Sanderson, Mark. Advantages of Query Biased Summaries in Information Retrieval. SIGIR 1998. Disponível em <http://portal.acm.org/citation.cfm?id=290947&coll=portal&dl=ACM&ret=1#Fulltext>

(Trivelpiece et al, 2000) Trivelpiece, A. et al. (2000) History of the Vision. In: Workshop Report on a Future Information Infrastructure for the Physical Sciences The Facts of the Matter: Finding, understanding, and using information about our physical world. Disponível em <http://www.osti.gov/physicalsciences/wkshprpt.pdf>

(Turtle & Croft, 1990) Turtle, H.; Croft, W. B. Inference networks for document retrieval. In Proceedings of the 13th International Conference on Research and Development in Information Retrieval, p. 1-24.

(Uitdenbogerd, 2000) Uitdenbogerd, Alexandra. Music IR: Past, Present and Future. 2000. MUSIC IR 2000. Resumo de Palestra Convidada. Disponível em http://ciir.cs.umass.edu/music2000/papers/invites/uitdenbogerd_invite.pdf

(van Rijsbergen, 1979) RIJSBERGEN, C. J. van. Information Retrieval. 1979. Disponível em <http://www.dcs.gla.ac.uk/Keith/Preface.html>.

(van Rijsbergen, 1986) van Rijsbergen, C. J. A new theoretical framework for information retrieval. In: SIGIR Conference Proceeding, p. 200. 1986.

(Vieira et al, 2000) Vieira, Renata; Gorziza, Fabiano; Rossi, Daniela; Chishman, Rove; Rossoni, Roberta; Pinnheiro, Clarissa. Extração de Sintagmas Nominais para o Processamento de Co-referência. Propor 2000. p. 19-22. Disponível em <http://www.inf.unisinos.br/~renata/>

(Vieira, 2001) Vieira, R. (2001) Resolução automática de correferência textual. I Congresso e IV Colóquio da Associação Latinoamericana de Estudos do Discurso ALED, Recife 23-28 de setembro.

(Wiley, 1998) Wiley, Deborah Lynne. Beyond Information Retrieval: Ways to Provide Content in Context. DATABASE 21, No. 4, p.18-22. 1998. Disponível em <http://www.onlineinc.com/database/DB1998/wiley8.html>.

(Williams, 2002) Williams, Robert V. "The Use of Punched Cards in US Libraries and Documentation Centers, 1936–1972". IEEE Annals of the History of Computing. Abril-Junho 2002. Vol 24, No 2. p. 16-33.

(Wu & Sonnenwald, 1999) Wu, Mei-Mei; Sonnenwald, Diane H.. Reflections in Information Retrieval Evaluation. Proceedings of the 1999 EBTI, ECAI, SEER & PNC Joint Meeting, 63-81. <http://pnclink.org/events-report/1999/Proceedings/wu-mm.pdf>

(Wurman, 1989) Wurman, Richard Saul. Information Anxiety. New York: Doubleday. 1989.

(Zipf, 1949) ZIPF, H.P., Human Behaviour and the Principle of Least Effort, Addison-Wesley, Cambridge, Massachusetts (1949).

Glossário

Acesso à Informação

É uma forma cuidadosamente construída de Recuperação de Informação. Seu objetivo é ajudar o usuário a descobrir, criar usos, reutilizar e entender a informação.

Conhecimento

É um conjunto de argumentos e explicações que interpretam um conjunto de informações. Trata-se de conceitos e argumentos lógicos essencialmente abstratos que interligam e dão significado a fatos concretos. Enquanto informação tem relação com a descrição, definição e perspectiva, conhecimento envolve estratégia, prática, método e metodologia. Informação indica o que, quem, quando e aonde. Já o conhecimento explica o como. O conhecimento é o que permite avaliar a informação de forma crítica e gerar nova informação.

Dados

São as evidências mais básicas, são os aspectos do fenômeno em estudo que podem ser captados e registrados. Correspondem a representações abstratas de observações diretas do mundo real, com relativamente pouca elaboração ou tratamento. Tais evidências, apesar de serem um reflexo razoavelmente confiável dos acontecimentos concretos, estão fora de contexto e portanto não tem relação nenhuma significativa com qualquer outra coisa.

Web escondida

É composta pelo conteúdo que está em bases de dados conectadas à web e que, portanto, só pode ser acessado através de consultas diretas. Quando requisitado, os resultados são dados através de páginas dinâmicas em tempo real. Apesar das páginas dinâmicas possuírem endereços (URLs) que as identificam de forma única, estes não são persistentes.

Descoberta de Conhecimento

Processo para extrair informações implícitas, novas e potencialmente úteis

encontradas em bases de dados grandes e heterogêneas e formular conhecimento. O objetivo do processo é analisar os dados sob diferentes perspectivas procurando por padrões e sumariá-los em informações úteis que possam ser utilizadas, por exemplo, para aumentar os lucros e/ou cortar custos. Segundo Fayyad (1997) este processo inclui: *“data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted knowledge”*.

Diretório

É um catálogo, que aparece para o usuário com uma página na Web, que possui um conjunto de categorias bem definidas. Em cada categoria existe uma coleção de links para páginas da Web que abordam assuntos relacionados com o contexto descrito pela categoria. Tais páginas em geral são categorizadas e revistas por editores humanos. Alguns dos diretórios mais conhecidos na atualidade são o Yahoo e o LookSmart encontrados respectivamente em: www.yahoo.com e www.looksmart.com.

Extração de Informação

O seu objetivo é analisar grandes volumes de textos para extrair tipos particulares de informação determinados por um conjunto de critérios de extração predefinidos. São extraídos fatos a respeito de eventos, entidades e relacionamentos preespecificados de documentos, que podem estar em línguas diferentes. Extração de informação retorna fatos para o usuário enquanto Recuperação de Informação retorna documentos.

Filtragem de Informação

O objetivo de um sistema de filtragem de informação é apresentar para o usuário apenas o que satisfaz suas necessidades dentre todo o volume de informação que tenha sido gerado. O processo de filtragem acontece apenas depois que já se tem acesso à informação. É aplicado a diversos domínios, como sistemas que chamam a atenção do usuário para novas informações e filtragem de notícias e e-mails. A diferença básica entre Filtragem e Recuperação de Informação é o fato de que enquanto a RI lida com a coleção e organização de textos e responde a interação do

usuário com os textos considerando apenas um episódio (uma única busca ou sessão) de busca, a filtragem lida apenas com a distribuição de textos para grupos e indivíduos e esta distribuição está também relacionada a mudanças entre diferentes episódios de busca. Ou seja, a filtragem é uma tarefa à parte que pode ser muito interessante para complementar/melhorar modelos de RI.

Hiponímia

É a relação que se estabelece entre duas palavras com base na maior especificidade de significado de uma delas, expressa relacionamentos é um tipo-de. Por exemplo: mesa está numa relação de hiponímia com móvel.

Informação

É o resultado de uma organização, transformação e/ou análise de dados. É o tratamento de um conjunto de dados de modo a produzir significado, de modo que possam então ser utilizados para dar suporte a decisões e outras ações.

Máquinas de busca

São sistemas de RI que aparecem para o usuário como uma página na Web e têm por objetivo encontrar informação de interesse dos usuários na Web. Coletam continuamente os dados disponíveis na Web e montam uma grande base de dados que é processada para aumentar a rapidez na recuperação de informação. Esta base de dados é coletada por robôs. As máquinas de busca também são chamadas em português de motores de busca, motores de procura e mecanismos de busca. Alguns exemplos são o google e o alltheweb, respectivamente encontrados em www.google.com e www.alltheweb.com.

Meronímia

Tipo de relacionamento é parte – de.

Meta ferramentas de busca

São sistemas de RI apresentados para os usuários como uma página na Web, mas ao contrário das máquinas de busca não constroem uma base de documentos. Submetem cada consulta a várias máquinas de busca, removem os resultados duplicados retornados pelas mesmas e sumarizam os resultados para o usuário. Dois exemplos são o metacrawler e o dogpile, encontrados respectivamente em:

www.metacrawler.com e www.dogpile.com.

Mineração de Dados

É o termo utilizado para referenciar a etapa do processo de descoberta de conhecimento em que são aplicadas técnicas/ferramentas para analisar e apresentar os dados, ou também como sinônimo de Descoberta de Conhecimento.

Mineração de Textos

Também procura por padrões, só que os procura em textos em LÍNGUA NATURAL (arrumar outros). O objetivo da mineração de Textos é analisar coleções de documentos como um todo para extrair informações que possam ser úteis para um determinado propósito. Tais informações podem ser esperadas ou não e podem também mostrar relacionamentos totalmente desconhecidos. Mineração de textos não é recuperação de informação, a recuperação de informação atende às necessidades de um usuário que foram expressas através de uma consulta retornando documentos, a mineração de texto explora relacionamentos entre documentos de forma independente das necessidades de um usuário.

Morfologia

É a parte da gramática que estuda a estrutura e a formação das palavras.

Nível de Coordenação

Número de termos que o documento tem em comum com a consulta.

Pergunta e resposta

É a tarefa de obter de grandes coleções de documentos respostas apropriadas para perguntas escritas em linguagem natural a respeito de um dado domínio. Esta área está altamente relacionada à extração de informação, recuperação de informação, interação em linguagem natural e outras áreas de pesquisa em PLN.

Prefixos

Cadeia de caracteres que inicia uma palavra. Por exemplo, o prefixo comput recupera palavras como computador e comutação.

Referenciación

A sucessão de coisas ditas ou escritas forma uma cadeia que vai além da

seqüencialidade: há um entrelaçamento significativo que aproxima as partes formadoras do texto falado ou escrito. Os mecanismos lingüísticos que estabelecem a conectividade e a retomada e garantem a coesão são os referentes textuais. Cada uma das coisas ditas estabelece relações de sentido e significado tanto com os elementos que a antecedem como com os que a sucedem, construindo uma cadeia textual significativa.

Robôs

Também chamados de *spiders*, *crawlers* ou *bots*, são programas que visitam cada página ou as páginas representativas de cada página da Web que deseja estar disponível para busca e as “lê” utilizando os hiperlinks para descobrir o endereço de outras páginas.

Sabedoria

É o resultado de entender os princípios fundamentais responsáveis pelos padrões que representam o conhecimento. A sabedoria tende a criar seu próprio contexto. Incorpora princípios, insights, lições e arquétipos. A sabedoria explica o porquê.

Sub-cadeia de caracteres

Cadeia de caracteres que pode aparecer em uma palavra. Por exemplo, tal está presente em tal, mortal, totalizado e talismã.

Tipologia

É o estudo dos diversos modos pelos quais as línguas podem diferir umas das outras.

Troponímia

Tipo de relacionamento similar à hiponímia, mas que não lida com substantivos ou sintagmas nominais e entidades, mas com verbos e processos (Green et al, 2002).