Uso de marcadores estilísticos para a busca na Web em português

Rachel Virgínia Xavier Aires

Uso de marcadores estilísticos para a busca na Web em português

Rachel Virgínia Xavier Aires

Orientadora: **Profa. Dra. Sandra Maria Aluísio**

Co-orientadora: **Dra. Diana Santos**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

USP – São Carlos
Agosto de 2005

Uso de marcadores estilísticos para a busca na Web em português

Rachel Virgínia Xavier Aires

Orientador: **Profa. Dra. Sandra Maria Aluísio**

Co-Orientador: **Profa. Dra. Diana Santos**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

"VERSÃO REVISADA APÓS A DEFESA"

Data da Defesa:	21/09/2005
Visto do Orientador:	

Este trabalho de doutorado foi desenvolvido por dois anos no NILC (Núcleo Interinstitucional de Lingüística Computacional) (www.nilc.icmc.usp.br) e por quase dois anos no pólo de Oslo da Linguateca (www.linguateca.pt), dois dos melhores centros de pesquisa em PLN que tratam da língua portuguesa. Tendo sido financiado pela Fundação para Computação Científica Nacional (FCCN), através da Fundação para a Ciência e Tecnologia e co-financiada pelo POSI (POSI/PLP/43931/2001) desde setembro de 2001.

Como crianças aos pés de um mágico, as pessoas reagem ao estilo por atitudes variáveis. Algumas simplesmente se descontraem e apreciam os efeitos, e talvez mais tarde recordem tranqüilamente suas emoções. Outras sentem o impulso do menino endiabrado, de espiar dentro da manga do mágico e expor as divisões do seu chapéu, mesmo com o risco de ignorar parte do seu espetáculo ou irritar o restante da audiência.

Nils Erik Enkvist (Enkvist *et al*, 1974)

Resumo

Como lidar com o excesso de informação ao qual usuários são submetidos em suas buscas na Web? São muitas as páginas sobre um mesmo assunto, por isso uma solução pode ser separá-las segundo os objetivos dos escritores. Melhor ainda seria separá-las segundo os objetivos dos leitores, tão diversos como buscar um programa, aprender sobre uma matéria ou saber as últimas notícias sobre um dado assunto. Esse é o objetivo desta tese, ir além do conteúdo dos textos para minimizar o esforço do usuário em encontrar os documentos que são relevantes para sua consulta em um dado instante de busca. Investigou-se pela primeira vez a hipótese de que é tecnicamente possível e de fácil compreensão a classificação resultados de busca segundo os seus objetivos. Para isso estudou-se a classificação automática dos resultados de buscas na Web em português segundo a intenção da busca. Foram aplicados algoritmos de aprendizado de máquina sobre características lingüísticas relacionadas com o estilo de documentos em português, e desenvolvidos estudos com usuários para avaliar na prática os classificadores criados. Foi também investigada a possibilidade de desenvolver classificadores personalizados que, dentro de um determinado assunto, separassem páginas interessantes de outras irrelevantes, com base em pequenos *corpora* de treinamento. Para a avaliação, foram utilizadas tanto as avaliações de sistema como as centradas no usuário. Os resultados mostram que (i) a classificação em necessidades é um conceito compreendido pelos usuários, (ii) o uso de marcadores estilísticos é um caminho barato e eficiente a ser investigado para obter classificadores confiáveis, (iii) o treinamento com pequenos *corpora* da Web é capaz de gerar classificadores confiáveis, e (iv) a busca pode ser facilitada por resultados classificados segundo necessidades de busca.

Abstract

How should one cope with information overflow, when there are too many pages on the Web about almost every subject? This thesis addresses the problem of information overflow users face when dealing with Web search results. To go beyond content, it is proposed to classify pages according to the search goals they serve from a user point of view: to download a system, learn some subject or find news about another are quite different user goals. The hypothesis validated in the present dissertation is that it is both technically feasible and understandble to classify Web pages according to user goal. By using machine learning techniques over linguistically inspired *features*, automatic classifiers were built to distinguish among user needs. Also, several user studies were conducted to assess the understandability of the concepts at stake and the gain achieved by using the particular classification in the display of the results. In addition, this work also tested personalized binary classifiers about specific subjects, trained in small training *corpora* supplied by the users themselves. With regard to evaluation, both system evaluation and user-centered evaluation were performed. The results show that (i) the user needs classification is understood by the user, (ii) the use of style markers are a reliable path to be investigated (iii) training on small Web *corpora* is able to generate reliable classifiers, and (iv) search can be eased by classifying search results according to user needs.

Agradecimentos

"A capacidade pouco vale sem oportunidade" Napoleão

Um dia eu pensei como Anthony Robbins, que disse "Engraçado, costumam dizer que tenho sorte. Só eu sei que quanto mais eu me preparo mais sorte eu tenho". Mas agora, no fim do doutorado, preciso admitir que o universo parece ter conspirado a meu favor. Muitos me ajudaram, minha família, minhas orientadoras, amigos, colegas, professores e mesmo totais desconhecidos. Graças a todos, nestes quase quatro anos aprendi muito. Vislumbrou-se o que quero para minha carreira. Conheci meu primeiro amor graças ao doutorado sanduíche na Noruega. Tive minha fé reforçada. Aprendi nas últimas semanas que não devo julgar antes do último segundo – o melhor é nunca julgar –, mas para isso ainda precisaria de uns mil doutorados. Tive provas freqüentes de que tudo acontece apenas no momento certo e porque realmente deveria acontecer. Obrigada a todos!

Obrigada às minhas orientadoras, Sandra e Diana, que cada uma ao seu modo, me orientaram e me "desorientaram". Obrigada, Sandra, pela amizade e paciência com meu mau humor dos últimos dias. Obrigada, Diana, pelas críticas diretas. Diana, quando começamos a trabalhar juntas, admirava sua capacidade de trabalho; hoje, além disso, respeito sua forma de trabalhar. Obrigada pelo apoio. Obrigada a ambas pela super dedicação nos últimos dois meses. Desculpem-me pelas inúmeras versões de textos de qualidade duvidosa a que as submeti.

Obrigada, família e amigos, por tolerarem minha ausência. Obrigada, meu querido amigo irmão Marcello, pelo zelo. Obrigada, *minino* Marcos pelo carinho a qualquer distância. Obrigada, Jorge e Tiago, por suas tentativas incansáveis de me tirar de casa aos finais-de-semana. Obrigada, Edvaldo, Alan, Claudete, Denise, Ludmila, Érika, Estela, Elma, Élide, Leandro, Ana Raquel, Flávia, Tulius e Taiz, por estarem em minha vida. Obrigada, Renatinho, Tânia, Lícia, Lilia, Helene, Cresita e Luís Costa, pelo apoio, foi muito bom conhecê-los melhor e conquistá-los para minha vida. Obrigada aos trabalhadores do grupo espírita Consciência e Caridade, pela boa energia e ajuda. Obrigada ao pessoal do NILC, pela companhia divertida. Obrigada, Tommy, pelo amor, dedicação e ouvido amigo. Obrigada também, amado, pela ajuda com as ilustrações e com o design do protótipo.

Obrigada a Akwan Information Technologies, pelos *logs* da máquina de busca TodoBr de novembro de 1999 e julho de 2002. Obrigada aos alunos e professores do ICMC que participaram do estudo apresentado em Aires & Aluísio (2003).

Obrigada, Aline, pelas conversas para portar os marcadores do Biber e pela ajuda para criar o *corpus* de necessidades. Obrigada, também a Crislaine, Vanessa e Lucélia, pela ajuda com a criação do *corpus*. Obrigada, Felipe, pela ajuda com o primeiro *script* de cálculo de *features*. Obrigada, Luiz, pelos *scripts* do Leva-e-traz. Obrigada, Leandro, e Gladis, por aplicarem o questionário do Apêndice B com seus alunos. Obrigada, Marcello, e Nana, por pedirem a seus colegas que também respondessem ao questionário. Obrigada, ao pessoal do NILC, Ariani, Lucas, Ricardo e Gawa, e aos amigos que participaram da avaliação final. Obrigada, Cristina, Nuno Cardoso, Luís Costa, Marcirio, Débora e Susana pelos corpora personalizados. Obrigada, a todos que dispensaram tempo para tirar minhas dúvidas e conversar sobre meu projeto.

Meus agradecimentos especiais a FCCN pelo apoio financeiro de setembro de 2001 a setembro de 2005.

Obrigada a todos!

ÍNDICE

Lista de Abreviaturas

ACM	*Association for Computing Machinery*
BNC	*British National Corpus*
FCCN	Fundação para Computação Científica Nacional
IRC	*Internet Relay Chat*
LMT	*Logistic Model Tree*
MEDLARS	*Medical Literature Analysis and Retrieval System*
NILC	Núcleo Interinstitucional de Lingüística Computacional
PLN	Processamento de Linguagem Natural
Propor	Processamento Computacional do Português Escrito e Falado
RI	Recuperação de informação
SIGIR	*Special Interest Group on Information Retrieval*
SMO	*Sequential minimal optimisation*
SOM	*Self-organizing map*
STASEL	*Stylistic Treatment at the sentence level*
SVM	*Support Vector Machine*
TREC	*Text Retrieval Conference*
WSJ	*Wall Street Journal*

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Publicações originadas da Tese

Aires, R.; Aluísio, A.; Santos, D. (2005) **User-aware page classification in a search engine.** Proceedings of 2005 SIGIR Workshop on Textual Stylistics in Information Access. SIGIR, agosto de 2005, Salvador – Brasil, 8 p.

Aires, R.; Santos, D.; Aluísio, A. (2005) **"Yes, user!": compiling a *corpus* according to what the user wants.** Corpus Linguistics 2005, julho de 2005, Birmingham – Inglaterra, 14 p. Disponível em www.corpus.bham.ac.uk/PCLC .

Aires, R.; Aluísio, S. (2005) "**As avaliações atuais de sistemas de busca na** Web **e a importância do usuário**". A ser publicado em Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.

Santos, D.; Simões, A.; Frankenberg-Garcia, A.; Pinto, A.; Barreiro, A.; Maia, B.; Mota, C.; Oliveira, D.; Bick, E.; Ranchhod, E.; Dias de Almeida, J. J.; Cabral, L.; Costa, L.; Sarmento, L.; Chaves, M.; Cardoso, N.; Rocha, P.; Aires, R.; Silva, R.; Vilela, R.; Afonso, S. (2004) **Linguateca: Um centro de recursos distribuído para o processamento computacional da língua portuguesa.** Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", p. 147-154, IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), novembro de 2004, Puebla - México.

Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) **Which classification algorithm works best with stylistic features of Portuguese in order to classify** Web **texts according to users' needs?** Relatório técnico nº 241, outubro de 2004, ICMC/USP.

Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) **What is my Style? Using Stylistic Features of Portuguese** Web **Texts to classify** Web **pages according to Users'Needs.** In LREC 2004, p. 1943-1946, maio de 2004, Lisboa - Portugal.

Aires, R.; Aluísio, S.; Quaresma, P.; Santos, D.; Silva, M. (2003). **An initial proposal for cooperative evaluation on information retrieval in Portuguese.** In PROPOR 2003 – 6th Workshop on Computational Processing of the Portuguese Language, p. 227-234, junho de 2003, Faro - Portugal. (c) Springer-Verlag.

Aires, R.; Aluísio, S. (2003). **Como incrementar a qualidade dos resultados das máquinas de busca: da análise de *logs* à interação em português.** Revista Ciência da Informação, vol 32, n. 1, p. 5-16, janeiro/abril de 2003.

Aires, R.; Santos, D. (2002). **Measuring the** Web **in Portuguese.** In EuroWeb 2002 conference, p. 198-199, dezembro de 2002, Oxford - UK.

Aires, R.; Aluísio, S. (2002). **Eu falo português. E daí?** In IHC 2002 – 5th Symposium on Human Factors in Computer Systems, outubro de 2002, Fortaleza - Brasil.

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)

Contextualização

Nos últimos anos, houve um crescimento explosivo do volume de informação. Livros, filmes, notícias, anúncios, música e, em particular, informações *on-line* surgem a todo o momento. Um estudo realizado na Universidade da Califórnia, em Berkeley, em 2000 (Lyman & Varian, 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 bilhão de gigabytes para ser armazenada. 1,5 bilhão que seria o equivalente a 250 megabytes por pessoa, isto é, para cada homem, mulher e criança na Terra. Para o ano de 2002, os mesmos autores acima, em um novo estudo (Lyman & Varian, 2003), estimam que tenham sido produzidos 5 exabytes de informação[1]. 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, dos quais de 10 a 20 terabytes seriam informação textual. 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-mail*s, listas de *e-mail*s, usenet e ftp. Já em 2003, o tamanho

[1] 5 exabytes seriam o equivalente a 37 mil bibliotecas com cerca de 17 milhões de livros cada (Lyman & Varian, 2003).

da Web foi estimado em 167 terabytes e da Web escondida estaria entre 66.800 e 91.850 terabytes (Lyman & Varian, 2003).

Um ponto interessante de se ressaltar sobre a Internet é o fato de ela ser uma mídia 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 & Varian, 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 negócios freqüentemente se sentem ultrapassados quanto a algumas mudanças no domínio em que trabalham. Mesmo enquanto pessoas comuns (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 poderiam gostar 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 essa 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 informações que não sabemos onde estão armazenadas, tentando nos atualizar e lendo informações que nunca serão utilizadas por nós. Todo esse esforço para gerenciar a informação acaba gerando custos extras para as organizações, tanto com armazenamento de informação como com pessoal. Além de gastos extras, uma conseqüência 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 e 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, ela está de alguma forma relacionada a áreas como descoberta de conhecimento (*knowledge discovery*), mineração de dados (*data mining*), mineração de textos (*text mining*), recuperação de informação (*information retrieval*) (RI), acesso à informação (*information access*), extração de informação (*information extraction*), resposta automática a perguntas (*question-answering*) e filtragem de informação (*information filtering*). O enfoque maior dessas 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 de acordo com nossa revisão bibliográfica pouco foi feito para garantir a qualidade da informação retornada.

Motivação e Relevância

Encontrar informação nesta nova mídia ou repositório de informação de tamanho imenso 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 entretenimento. O conteúdo disponível na Internet aumenta constantemente e as pessoas que incluem novos dados, em sua maioria, não sabem

como 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 & Varian, 2000). Em novembro de 2002, Aires & Santos (2002) estimaram o tamanho da Web em português em 20.807.956 páginas no Alltheweb, 7.152.022 páginas no Altavista e 4.260.000 páginas no Google. Em junho de 2005, o mesmo experimento foi replicado e encontramos um número de páginas bem superior ao encontrado em 2002, que foram, respectivamente, 149.000.000, 167.000.000 e 19.100.000 páginas.

De acordo com nossa pesquisa bibliográfica são diversos profissionais trabalhando na Recuperação de Informação no Brasil e em Portugal sob diferentes perspectivas como, por exemplo, psicólogos, bibliotecários, pesquisadores da área 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. Essa é uma das razões da 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.

Objetivos

O objetivo inicial desta tese era investigar a utilização do Processamento de Linguagem Natural (PLN) na RI na Web em português. Por um lado, aplicar técnicas de PLN do português à RI, por outro, lançar alguma luz sobre as características, que supomos ser diferentes, da Web brasileira e dos usuários brasileiros e/ou em português.

Esta é a primeira tese que parte dos problemas concretos dos usuários em português em vez de simplesmente aplicar técnicas já desenvolvidas para o inglês ou para a Web em geral.

Após alguns estudos preliminares de como obter os objetivos de consultas (Aires & Aluísio, 2003), de ter uma noção estatística das consultas reais (Aires *et al*, 2004b) e de ter definido o problema de uma forma global (Aires *et al*, 2003), escolhemos nos dedicar à melhoria da parte da RI mais diretamente relacionada com os usuários: a apresentação dos resultados. Decidimos implementar um meta-buscador, e estudar a categorização das respostas em esquemas de classificação que fossem compreensíveis e úteis aos usuários. Assim, o objetivo principal desta tese é estudar, para o português, que categorizações dos textos e páginas da Web permitem uma forma mais fácil de organização dos resultados de uma busca, e como obter automaticamente essa categorização.

Para esse objetivo, investigou-se o uso de características estilísticas e de um *corpus* de páginas Web classificadas segundo as necessidades que satisfazem.

Além disso, foram realizados alguns dos primeiros experimentos com usabilidade associados à busca na Web em português, estudos esses que esperamos ser motivadores e um bom ponto de partida para outros pesquisadores, assim como toda a abordagem e preocupação com a avaliação que tivemos.

Organização da Tese

Esta tese está dividida em três partes com onze capítulos no total. A primeira parte trata do processo de recuperação de informação textual e apresenta uma breve história sobre a evolução da recuperação de informação nos últimos 50 anos (Capítulo 1); descreve as principais formas de avaliação de sistemas de RI encontradas na literatura (Capítulo 2) e discute como os sistemas de recuperação de informação fazem e poderiam fazer uso de recursos e técnicas de PLN na tentativa de aumentar sua precisão e revocação (Capítulo 3). A segunda parte detalha a hipótese deste trabalho (Capítulo 4); define estilo e marcadores estilísticos e exemplifica o uso dos mesmos para solucionar diferentes problemas (Capítulo 5); define os esquemas de classificação escolhidos, os *corpora*, algoritmos e conjuntos de marcadores estilísticos utilizados e o protótipo de um meta-buscador desenvolvido (Leva-e-traz) (Capítulo 6). A última parte mostra quatro avaliações com objetivos diferentes: verificar que esquemas de classificação interessam aos usuários (Capítulo 7); avaliar os

classificadores desenvolvidos sob o ponto de vista do sistema de busca (Capítulo 8); avaliar com usuários se a proposta de uso de necessidades personalizadas é bem interpretada e como os marcadores estilísticos funcionam para problemas reais (Capítulo 9) e verificar com uma avaliação do protótipo Leva-e-traz sob o ponto de vista do usuário se a classificação em necessidades auxilia o usuário a determinar quais resultados de busca realmente atendem às suas necessidades (Capítulo 10).

I

Recuperação de Informação tradicional e com PLN

1. Recuperação de Informação

"The way we see the problem is the problem." Stephen R. Covey

Recuperação de Informação (RI) (*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 a essa necessidade. Por itens de informação entende-se informação em diferentes mídias, tais como: textos, imagens (fotografias e mapas), vídeos. 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 (A*udio Information Retrieval*) (Uitdenbogerd, 2000);
- RI multimídia (*Multimedia Information Retrieval*) (Chiaramella *et al*, 1996).

A RI textual 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 monolíngüe é a habilidade do primeiro 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 1.1, enfatizando a linguagem de consulta, a de indexação e modelos de recuperação. Concluímos o capítulo com uma breve discussão sobre a evolução das técnicas de RI desde a década de 40 até os dias atuais (Seção 1.2).

1.1 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, responsáveis por: 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 à necessidade do usuário.

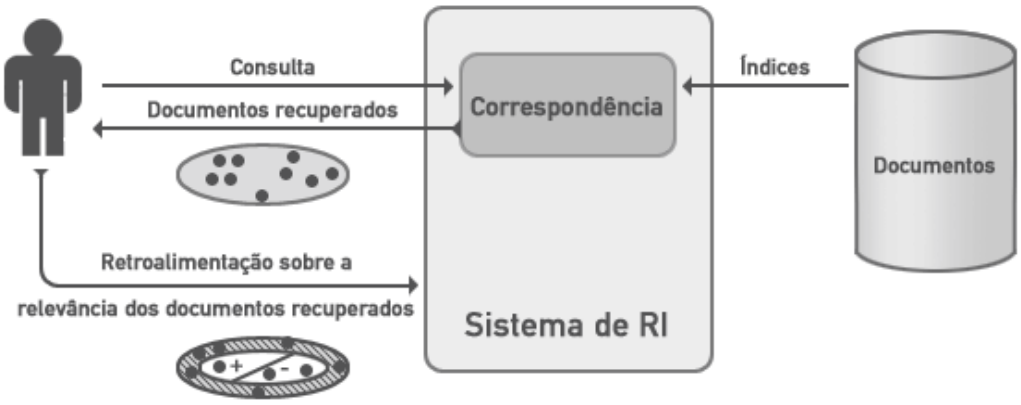

Figura 1 - Processo típico de RI (Belew, 2000)

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 todo o conteúdo do 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 dessa classe.

Luhn[2] (1958) propõe que a freqüência seja utilizada para extrair palavras e sentenças representativas de um documento. Dada uma freqüência f de ocorrência e a ordem r (*rank*) dessa freqüência de ocorrência, então um gráfico relacionando f a r seria uma curva similar à mostrada na Figura 2, que diz que o produto da freqüência

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

de uso de uma palavra e sua ordem de importância é aproximadamente constante. Luhn utiliza esta lei, a 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.

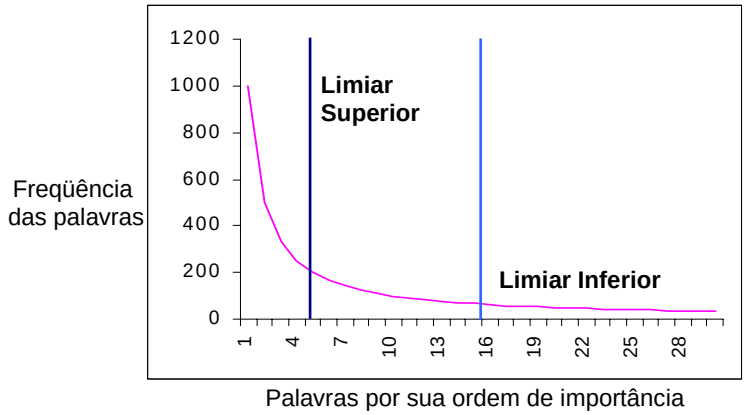

Figura 2 - Lei de Zipf (Figura adaptada de van Rijsbergen, 1979, p.16, Figura 2.1)

A remoção das palavras de alta freqüência (*stopwords*)[3] é uma forma de implementar o limiar superior – isso pode ser feito comparando a entrada com uma lista de *stopwords*. Um passo complementar seria remover sufixos (*suffix stripping*), assim muitas palavras equivalentes seriam mapeadas em uma única forma. Outro passo seria checar os radicais, supondo que se duas palavras possuem o mesmo radical (*stem*), essas 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 significante 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 1.1.2 descreve as principais técnicas para construção de arquivos de indexação.

[3] Veja exemplos de listas de *stopwords* para 17 línguas em http://www.ranks.nl/*stopwords*/ .

Para aumentar a chance de se obter documentos relevantes, pode-se ainda contar com a ajuda de um tesauro (Jing & Croft, 1994), o que pode ser feito substituindo-se 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 1.1.1. 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 para a 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-se cada consulta aos documentos armazenados ou aos *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 consultas, e como esses são comparados (Turtle & Croft, 1990). Alguns exemplos de modelos são: Modelo Booleano (Paice, 1984), Modelo de Espaço Vetorial (Salton & McGill, 1983), Modelo Probabilístico (Maron & Kuhns, 1960), Modelos Booleanos Estendidos (*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 da Língua (*Statistical Language Models/Language Models*)

(Ponte & Croft, 1998). Na Seção 1.1.3 explicamos os modelos clássicos de recuperação, isto é, booleano, vetorial e probabilístico.

A **saída** do sistema de RI 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, resumo, 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.[4]

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 para 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 (em 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 documentos visitados ou gerando novas consultas com o uso de tesauros e/ou ontologias.

1.1.1 Linguagem de consulta

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

[4] www.kartoo.com

As **consultas através de palavra-chave** são o tipo comumente aceito por sistemas de RI. Podem ser compostas somente por palavras soltas e, nesse caso, o resultado retornado pelo sistema é um conjunto de documentos que contêm pelo menos uma das palavras da consulta, ordenados pela freqüência das palavras nos documentos (*term frequency*) ou pela freqüência inversa (*inverse document frequency*). As consultas por palavras soltas podem ainda ser consideradas dentro de um contexto, procurando-se 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 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 como frases de uma 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. Isso 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 ela seja representada para o sistema com várias palavras ou frases. Nesse 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,[5] sufixos, subcadeias

[5] Por exemplo, o prefixo “comput” recupera palavras como “computador” e “comutação”.

de caracteres, intervalos (*ranges*), palavras semelhantes, expressões regulares e padrões estendidos.

Intervalos (*ranges*) são utilizados para cobrir quaisquer palavras que estejam entre um par de cadeias de caracteres seguindo a ordem alfabética, por exemplo, o intervalo 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 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 seqüência que confira, por exemplo, com 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-mail*s, cada um com os campos: remetente, data, assunto e corpo de texto, é possível procurar *e-mail*s enviados por uma determinada pessoa com a palavra "avaliação" no campo assunto.

1.1.2 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, essa lista é dividida em dois arquivos: de vocabulário e de endereçamento. O arquivo de vocabulário contém 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 é o alto custo para atualização do índice (Frakes & Baeza-Yates, 1992).

Arquivos de Assinatura são estruturas de indexação orientadas por palavra baseadas em *hashing*; são compostos por vários blocos de assinatura (Kowalski, 1997). 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* que determina quais posições da assinatura devem ser setadas para 1. Depois de determinadas as assinaturas de todas as palavras de um bloco, elas são combinadas (geralmente por uma função estilo OR) a fim de criar a assinatura do bloco.

Os documentos são divididos em blocos lógicos contendo, cada um, um número n de palavras, a fim de evitar que as assinaturas sejam muito densas (o que ocasionaria uma grande quantidade de colisões, ou seja, palavras com assinaturas similares). Quanto maior for a assinatura, menor é a possibilidade de colisões. São

apropriados para textos que não sejam muito longos; na maioria das aplicações os arquivos invertidos possuem uma performance superior à dos arquivos de assinatura (Frakes & Baeza-Yates, 1992). Um exemplo pode ser visto na Figura 4 (fonte Kowalski, 1997), na qual o tamanho do bloco é de 5 palavras, o tamanho da assinatura é de 16 bits e o número máximo de dígitos "1" permitidos é 5.

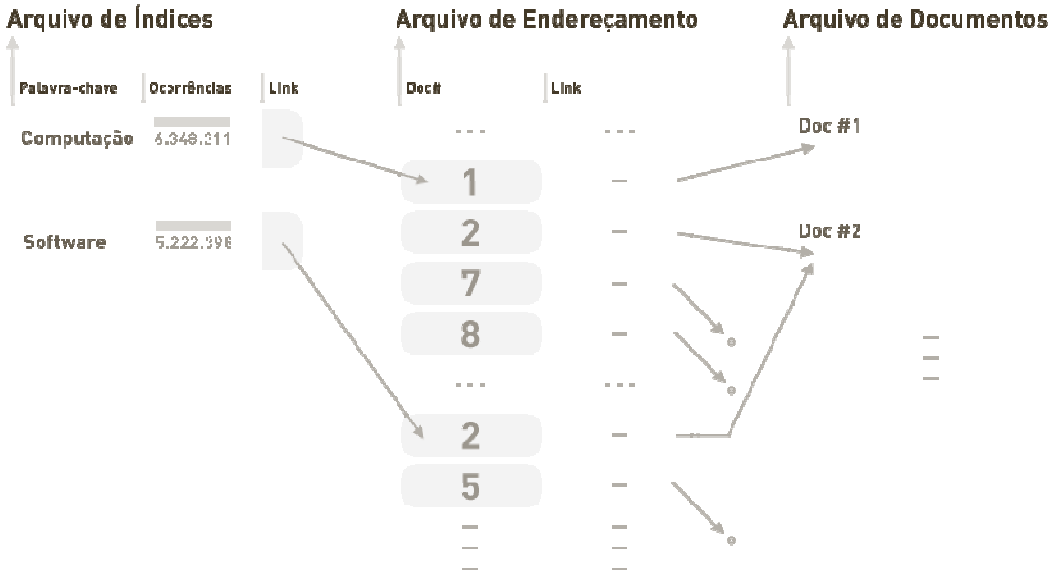

Figura 3 - Exemplo de Arquivo ou Índice Invertido

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 bloco:	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). A Figura 5 mostra a árvore

de sufixos para a *string* “xabxac”, cujos sufixos são “xabxac, abxac, bxac, xac, ac, c”. Uma boa introdução sobre árvores de sufixos é dada por Gusfield (1997).

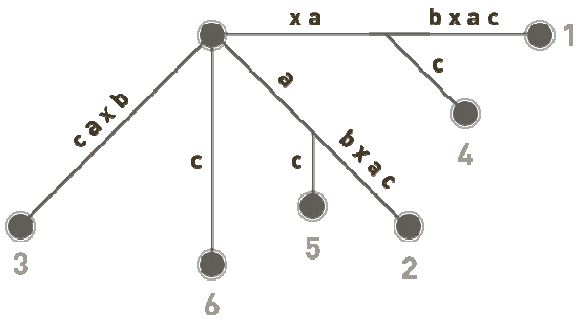

Figura 5- Exemplo de árvore de sufixos

1.1.3 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 e Probabilístico.

1.1.3.1 Modelo Booleano

O modelo booleano foi o primeiro modelo utilizado em RI e o mais utilizado até meados da década de 1990, apesar das alternativas de modelo que surgiram desde o final dos anos 1960.

O modelo booleano considera uma consulta como uma expressão booleana convencional, que liga seus termos através de conectivos lógicos AND, OR e NOT. Nesse 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 de o 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.[6]

Outra característica desse 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 uma forma de controlar/direcionar 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".

1.1.3.2 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),..., (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.

[6] Alguns exemplos são Google e AlltheWeb, respectivamente encontradas em www.google.com e www.alltheweb.com.

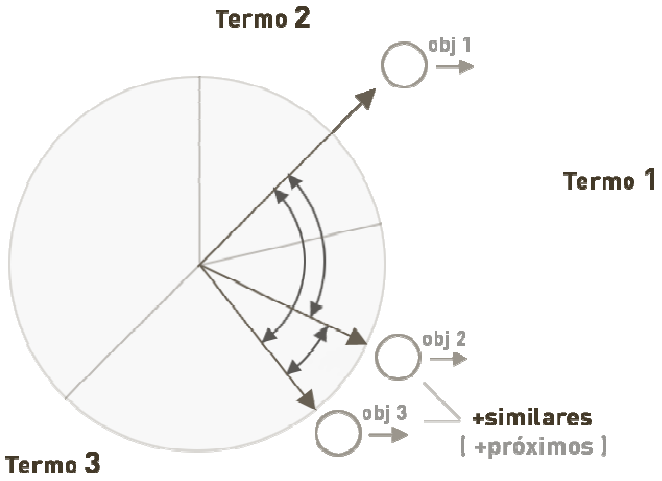

Figura 6 - Similaridade de documentos no modelo vetorial

Consultas também são representadas por vetores. Dessa 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:

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

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 os termos no modelo, para modelar, por exemplo, frases ou termos que aparecem perto um do outro. Esse modelo traz ainda duas dificuldades: a associação de pesos aos termos, que nem sempre é uma tarefa simples, e a implementação propriamente dita.

1.1.3.3 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*). Nesse 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
- $-R_a$ — 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

Dada uma consulta Q_a, o modelo probabilístico atribui a cada documento D (como medida de similaridade) um peso $W_{D/Qa}$, 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: $D=\{x_1, x_2, ..., x_n\}$, $x_k \in \{0,1\}$. 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 a 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:

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

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:

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

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

$$W_{D/Q_a} = \sum_{k=1}^{n} x_k \times w_{ak} + C$$

$$x_k \in \{0,1\}$$

$$w_{ak} = log\frac{r_{ak}}{1-r_{ak}} + log\frac{1-s_{ak}}{s_{ak}}$$

$$C = log\frac{P(+R_a)}{P(-R_a)} + \sum_{k=1}^{n} log\frac{1-r_{ak}}{1-s_{ak}}$$

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:

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

onde:

- $W_{D/Qa}$ é a medida de similaridade entre a consulta Qa 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 desse 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.

1.2 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 1940 (Luhn, 1959; Ohlman, 1998) e durante a década de 1950 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. Mooers.[7] Os sistemas desta primeira geração de sistemas de RI eram compostos basicamente por catálogo de cartões (Williams, 2002), contendo, em geral, o nome do autor e o título do documento.

A década de 60 foi uma época de muitos experimentos em RI. As métricas precisão (*precision*) e revocação (*recall*) usadas na área de 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 1960 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 resposta automática a perguntas. Foram várias as publicações na década de 1960, relacionadas dentre outros tópicos a modelos probabilísticos, sistemas booleanos e ao modelo vetorial, por exemplo: Maron & Kuhns (1960), Becker & Hayes (1963), Sparck Jones (1964), Salton (1968). Foi desenvolvido o sistema MEDLARS[8] (*Medical Literature Analysis and Retrieval System*), o primeiro grande

[7] http://www.tracfoundation.org/mooers/mooers.htm
[8] Para uma descrição atualizada de MEDLARS, veja Parris (1998).

sistema de RI a utilizar uma base de dados informatizada e o processamento de consultas em *batch*.

Na década de 1970, surgiram os primeiros processadores de texto e muitos textos começaram a ficar disponíveis em formato eletrônico. Foram desenvolvidos os primeiros sistemas *time-sharing* – as consultas passaram a ser apresentadas diretamente em terminais e o usuário pôde ter a resposta imediatamente. Alguns exemplos são: NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT. Foi nesta década também que se intensificaram as pesquisas em RI Probabilística. Alguns exemplos de publicações referentes a avanços teóricos e métodos de atribuição de pesos estatísticos da década de 1970 são: Jardine & van Rijsbergen (1971), Salton (1975), Salton *et al* (1975a, 1975b), Van Rijsbergen (1979). Outro acontecimento importante desta década de 1970 se deu em 1978 com a primeira conferência da *Association for Computing Machinery* (ACM) dedicada à Recuperação da Informação — *Special Interest Group on Information Retrieval* (SIGIR[9]).

Na década de 1980, o processamento de textos continuou a crescer e o preço do espaço em disco começou a cair. Com isto e também com o desenvolvimento do CD-ROM, muito mais informações (textos completos) ficaram disponíveis, inclusive informações não textuais, o que fez despertar um interesse ainda maior pela RI multimídia. Porém, os avanços e novas direções da pesquisa nessa área, como, por exemplo, a preocupação por técnicas de indexação capazes de lidar com grandes volumes de dados rapidamente, só foram efetivamente vistos na década de 1990. Foi na década de 1980 que os sistemas de RI passaram também a ser utilizados por não especialistas e que muitas das técnicas desenvolvidas anteriormente passaram a ser realmente aplicadas.

Na década de 1990, a comunidade de RI viu as tecnologias saírem da fase experimental para a fase de uso e serem amplamente testadas devido à velocidade com a qual foram adotadas durante as décadas de 1970 e 1980 pelas aplicações comerciais. Foi em 1992, por exemplo, que aconteceu pela primeira vez a conferência sobre Recuperação de Informação Textual – *Text Retrieval Conference* (TREC).[10]

[9] http://www.acm.org/sigir/
[10] http://trec.nist.gov/

Nos anos 1990 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 começou 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 pudesse integrar bem os sistemas de RI com os outros sistemas de uma organização.

Foi também no final dos anos 1990 que surgiram 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. Isso 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 às características dos dados encontrados na Web (distribuídos, voláteis, em grande volume, não estruturados, nem sempre de boa qualidade e heterogêneos) e ao perfil dos usuários desses sistemas. Na Tabela 1, mostramos como algumas das técnicas da RI discutidas nas seções anteriores foram adotadas por ferramentas de busca.

Diferentemente 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 a coleta de documentos, para que se pudesse cobrir uma grande porcentagem da Web e se manter uma coleção atualizada (Belew, 2000). Além disso, 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 s*pam* (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,[11] por exemplo, possibilita o uso de um filtro para omitir o que eles chamaram de conteúdo ofensivo, e a máquina de busca Google[12] 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.

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

A outra principal diferença entre os sistemas convencionais e os sistemas na Web que causou mudança nas pesquisas nos anos 1990 foram os usuários. Na Web, não existe um usuário típico, há usuários totalmente inexperientes e usuários com os

[11] www.alltheWeb.com
[12] www.google.com

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 a 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 que o ajudem a entender a origem dos erros;

(ii) interfaces simples para usuários não especialistas; e

(iii) interpretação de consultas ambíguas, sem objetivo claro, ou simplesmente que contenham erros de digitação ou de ortografia.

Nos últimos anos, continuam a merecer atenção os tópicos considerados na década de 1990. 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 nessa 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 desse tipo de pesquisa, já que os usuários desse tipo de sistema procuram informações para as utilizarem logo após a consulta no local onde se encontram, por exemplo, procurando pelo telefone da oficina mais próxima ao local onde 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 Artificial 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 esses que se não tratados, ainda que parcialmente, impossibilitarão maiores avanços na melhoria da

precisão e revocação dos sistemas de RI textual. No Capítulo 3 apresentamos alguns esforços que foram feitos nesse sentido com uso de técnicas/recursos de PLN e de características da língua e também apresentamos uma visão resumida do que o PLN ainda pode oferecer. No próximo capítulo apresentamos um breve resumo sobre a avaliação de sistemas de RI.

2. Avaliação de sistemas de RI

"Usa a estatística como o bêbado utiliza o poste, mais pelo apoio do que pela luz." Autor desconhecido

Sistemas de Recuperação de Informação têm sido avaliados e comparados há vários anos. Na década de 1960, 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; (ii) 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 2.2. As avaliações que são baseadas em julgamentos de relevância são as chamadas **avaliações centradas no sistema**; **outra alternativa são as avaliações centradas no usuário**. Essas duas abordagens são descritas na Seção 2.1, onde mostramos algumas de suas vantagens e desvantagens. Na Seção 2.3 mostramos algumas das medidas baseadas em relevância mais utilizadas e também algumas das revisões feitas sobre essas medidas para que elas fossem utilizadas na avaliação de sistemas de RI na Web. Na Seção 2.4 apresentamos os detalhes que devem ser considerados na criação de um conjunto de teste para uma avaliação baseada em relevância.

2.1 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. Essas avaliações não seguem uma metodologia padrão de avaliação, fazem uso de técnicas e medidas de avaliação de outras áreas como, por exemplo, da área de Interação Usuário-Computador 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 de a medição da eficácia ser o foco principal dessas avaliações é a hipótese de que quanto mais eficaz for um sistema, mais ele atenderá às necessidades do usuário. Diferentemente das avaliações centradas no usuário, que são feitas através de experimentos interativos, as centradas no sistema utilizam um conjunto de testes, que é composto por uma coleção de documentos, uma lista de consultas/requisições e de julgamentos de relevância. Exemplos de conjuntos de testes são os utilizados na avaliação conjunta da *Text Retrieval Conference* (TREC).[13]

Existem várias críticas às avaliações centradas no sistema:

- por serem realizadas em ambientes de "laboratório" e não em ambientes reais;
- pelo problema de credibilidade dos julgamentos de relevância, já que este é um conceito subjetivo (Wu & Sonnenwald, 1999);
- pelo problema de representatividade do conjunto de consultas e de documentos, uma vez que costumam ser voltados para o domínio da ciência e tecnologia. Em contrapartida, esse 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.

2.2 Relevância

A noção de relevância é **subjetiva**, pois 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

[13] http://trec.nist.gov/data.html

da área. Já para um outro usuário aluno de graduação procurando 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**, dado que o mesmo usuário pode fornecer julgamentos de relevância diferentes para os mesmos documentos e consultas com a mudança do contexto de uso da informação. Por exemplo, uma professora quando prepara 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 artigos sobre o mesmo tema para se inteirar das novidades, porque irá participar de uma banca de doutorado, os artigos avançados serão altamente relevantes.

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 o é em relação ao primeiro. Tentar manter todos os julgamentos consistentes no caso desse exemplo seria uma tarefa árdua.

Relevância é também **dinâmica**. Uma consulta pode não ser totalmente satisfeita por um único resultado, pois 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 esse 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 ela não foi totalmente atendida, passar a considerar os resultados que foram a princípio considerados irrelevantes como relevantes. Por exemplo, se uma 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, esses 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 documentos; tem-se acesso apenas a partes dos documentos: título, primeiro parágrafo e citações bibliográficas. Nesse caso, o julgamento irá ainda depender de quão informativos são os dados fornecidos sobre o documento.

2.3 Revocação, precisão e outras medidas de eficácia

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. Suas fórmulas são mostradas abaixo.

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

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

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

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

A relação entre essas 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 compõem 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 a um sistema podemos calcular a sua 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[14] (*coordination level*), o par precisão/revocação pode ser calculado 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 ilustrado no Gráfico 2.

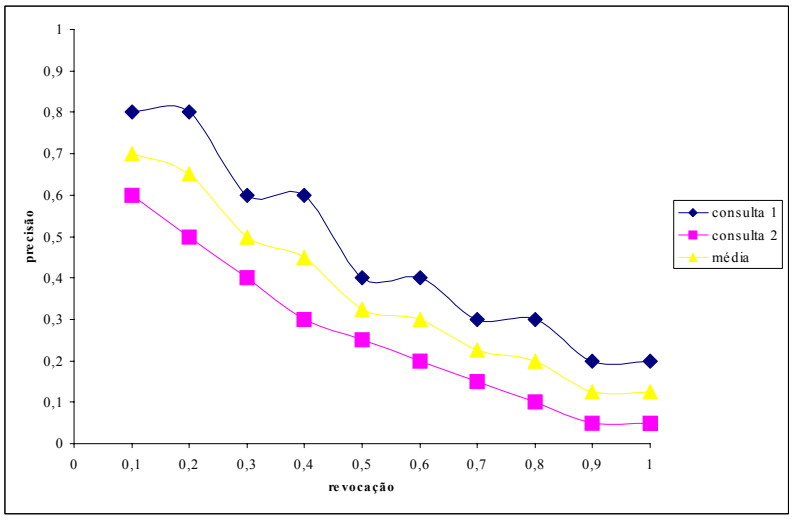

Gráfico 1 – Curvas de precisão/revocação

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 independentemente de qualquer parâmetro. Nesse último caso, uma forma de se montar a curva precisão-revocação do sistema é utilizando-se vários

[14] O Nível de coordenação indica o número de termos que o documento tem em comum com a consulta.

valores de revocação pré-estabelecidos. Nesse caso, é calculada a média das precisões (precisão média) para cada curva para cada um dos valores de revocação pré-estabelecido.

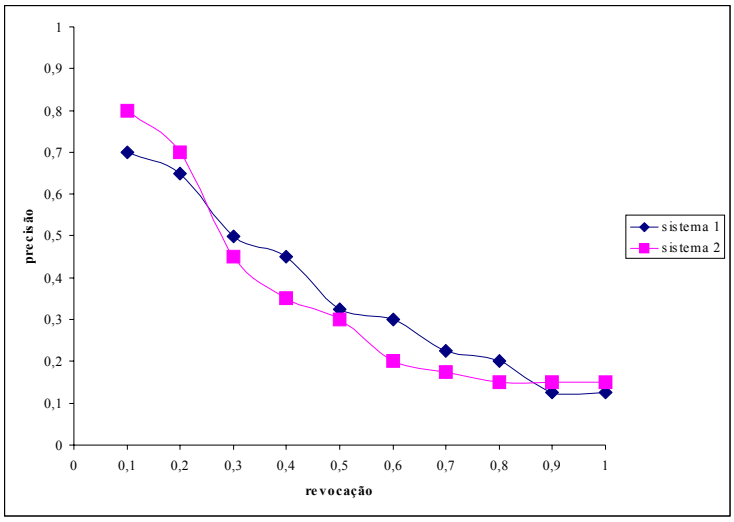

Gráfico 2 - Comparação de sistemas utilizando-se as curvas de precisão/revocação

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 dessas 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 β 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 $\beta > 1$, o peso da precisão é maior que o da revocação, e quando $\beta < 1$, 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).

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

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

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

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.

2.3.1 Medidas influenciadas por características da Web

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 de RI na Web, é de se

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 e o de 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 esse 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 desse exemplo, a revocação é a medida mais importante, para o aluno é a precisão. No semestre seguinte esse 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.

Como conclusão de nossa revisão bibliográfica, 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.

(2) Um documento que não responde completamente à pergunta, mas tem *links* para documentos que respondem pode ser considerado um documento relevante.

(3) Muitos usuários não conferem todos os primeiros 10 ou 20 resultados quando não encontram nenhum documento relevante entre os primeiros listados.

(4) Itens duplicados ou que não levam a nenhum documento não são relevantes.

Algumas avaliações já consideram algumas das características acima. 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 2.

Tabela 2- 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).

2.4 O conjunto de teste

São duas as formas utilizadas, em geral, para criar um 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. Nesse 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. Nesse caso, a base de documentos será formada pelos resultados das consultas. A listas das consultas, nesse 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 de um conjunto de teste, no caso de a 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. 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 se 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, informando se a veracidade da informação é importante, se basta que o resultado aborde 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.

3. RI e Processamento de Linguagem Natural

"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)

Como vimos no Capítulo 1, os sistemas de RI tradicionais se baseiam basicamente em estatística e matemática. Consideram as palavras de um documento (e as da consulta) como unidades atômicas isoladas, utilizando principalmente a freqüência desses termos no texto. Os métodos estatísticos foram utilizados na RI a partir da década de 1960 após constatação da dificuldade em se utilizar as técnicas de PLN na RI. Entretanto, realizar a indexação de documentos e a busca utilizando somente métodos estatísticos compromete a eficácia, fato percebido desde o fim da década de 1980 (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 língua do texto (Smeaton, 1991, p. 374, *apud* Bräscher, 1999) mesmo que o preço a ser pago seja o custo do processamento da língua natural. Entretanto, esse 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 freqüência 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 na criação de índices; interpretação de consultas e retroalimentação; comparação entre consultas e índice; apresentação de resultados e diálogo, respectivamente. Concluímos o capítulo (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 ele estar escrito em uma língua desconhecida ou diferente da utilizada na consulta, não apresentaremos as técnicas que são utilizadas nesses tipos de sistemas, por exemplo, os recursos de tradução e indexação semântica latente (Deerwester *et al*, 1990). 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 conhecimento 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 freqüência 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-mole"[15] 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, referimo-nos 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 é mais defendido na RI são os constituídos por sintagmas nominais (Arampatzis *et al*, 1999).

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

[15] 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.

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 lexicais, 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. Essa 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. Essa 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 modificados facilmente. 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 geram 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, ou seja, 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, ela é feita através do uso de um *stemmer* ou da 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 tesauros, 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*).

A Indexação Semântica Latente se baseia na observação de que uma matriz de termos de índices por documentos é esparsa, pois a maioria dos termos não aparece na maioria dos documentos, culminando numa matriz com muitos valores nulos. Essa 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, os significados/sentidos de palavras e termos, pois 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 o 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 que utilizam *stemmers* ou que utilizam 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 do 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 o 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 isso, no cálculo do grau de certeza consideram a certeza para o radical e a certeza para o sufixo. No Quadro 1 mostramos um resumo de algumas das técnicas e recursos de PLN que podem ser utilizados na indexação.

Quadro 1 – PLN para melhoria do Indexador

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 *Stemmer* Tesauros

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.

Vários autores (Abdulla *et al*, 1997; Cacheda & Viña, 2001a; Spink *et al*, 2002) salientaram que as consultas são geralmente pequenas, muitas com apenas duas ou três palavras. 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 do modo como 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 de esse tipo de entrada ser aparentemente melhor para usuários pouco experientes, que são o tipo de usuário cada vez mais comum para os novos de sistemas de RI, ele não é o mais adequado para sistemas de RI em meios como telefones celulares, já que nesses 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 necessária a geração de novas consultas.

No caso de a consulta ter sido mal formulada por conter erros ortográficos, o sistema pode procurar palavras semelhantes às utilizadas na consulta para modificá-la automaticamente ou apresentar outras palavras como alternativas para o usuário. Isso pode ser feito, por exemplo, utilizando-se *stemmers*, dicionários ou técnicas mais avançadas de correção ortográfica.

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, isto é, da 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 nesse processo de refinamento. Esse processo pode ser feito com ou sem a interferência do usuário. Podem ser utilizados os documentos que o usuário tenha considerado 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 esses aparentemente entendem como a retroalimentação funciona e encaram a retroalimentação como uma forma de ter mais facilmente controle sobre o sistema (Koenemann & Belkin, 1996). A retroalimentação é utilizada atualmente em várias implementações (Robertson & Sparck Jones, 1996).

Uma outra opção para se refinar a consulta 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 tanto 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 contenham 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, 2001b, 2001c). 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 homonímia 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. Sanderson (1994), por exemplo, afirma que a aplicação de técnicas de desambiguação de significado na RI não resultou em resultados mais precisos.

Como vimos no Capítulo 1, 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. Nesse 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 o 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).

No Quadro 2 mostramos um resumo de algumas das técnicas e recursos de PLN que podem ser utilizados na interpretação de consultas.

Quadro 2 – PLN para a interpretação das consultas

Intérprete de consultas	
Questão tratada	O que o usuário realmente deseja
Tarefas	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 morfossintática Análise morfológica Análise sintática Corretor ortográfico Extração de múltiplos termos Extração de relações semânticas Tesauros e redes semânticas Uso de padrões de perguntas em língua portuguesa

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, pois fazem uso constante de sistemas *on-line*, a estratégia de utilizar algoritmos mais simples, mas que não interferem na performance, na comparação de índices e consultas é uma idéia mais do que interessante para os sistemas atuais, não só para as ferramentas de busca. Por isso, nesta seção tratamos de duas técnicas que aparentemente não aumentariam a complexidade do mecanismo de comparação: o uso de textos segmentados e de características do estilo do texto.

Essas técnicas não estão diretamente relacionadas ao assunto de um documento, mas podem ser utilizadas para ajudar a encontrar quais são os documentos realmente relevantes para um determinado usuário. Razão disso 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 Características estilísticas de um texto

Variações estilísticas são tão freqüentes em textos sobre um mesmo assunto quanto a variação de assuntos entre textos do mesmo gênero ou variedade (Karlgren, 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. Estilo é tratado em maiores detalhes no Capítulo 5.

No Quadro 3 mostramos um resumo de algumas das técnicas e recursos de PLN que podem ser utilizados na correspondência entre documentos e consultas para escolha dos resultados a serem retornados.

Quadro 3 – PLN para a 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	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

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 considera 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. Esse tipo de apresentação dos resultados 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. Essa visão facilitaria o uso do sistema nas novas consultas do usuário e até mesmo o refinamento da última consulta feita por ele. Para informar o usuário sobre como o sistema funciona, podemos, por exemplo, ordenar os resultados agrupando-os pelas organizações ou autores que produziram os documentos. 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 tratados nos documentos.

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 tipos 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 esse 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. Sumário esse que, quanto mais estiver voltado para a consulta do usuário, mais facilitará 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 esses últimos melhoraram tanto a eficácia 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 desses 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 ser mais apropriada para aquele padrão ou até mesmo para 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 Considerações sobre RI e PLN

Ao levantar as melhorias que poderiam ser feitas em cada uma das fases da RI, encontramos vários autores defendendo a importância do uso de técnicas lingüisticamente motivadas na RI, em especial, o uso de técnicas e recursos de PLN em diferentes fases da RI. Esses usos iriam desde análises mais profundas na criação de índices a análises superficiais na apresentação dos resultados. Um resumo das técnicas, recursos e pesquisas que poderiam ser utilizados para realizar as melhorias citadas nas quatro sessões anteriores é mostrado no Quadro 4.

Quadro 4 - Técnicas, recursos e pesquisas que podem melhorar a qualidade dos sistemas de RI

	Índice	Consulta	Comparação	Resultados	Retroalimentação
Análise de tema de um texto	■				
Análise morfológica	■	■			
Análise sintática	■	■			
Classificação de textos quanto ao gênero			■	■	
Colocações	■	■			
Co-referência	■				
Estudos sobre a influência do tamanho dos documentos na qualidade do conteúdo	■		■	■	
Ontologias	■	■	■	■	
Pergunta-resposta			■	■	
Reconhecimento de nomes	■	■		■	
Segmentação de textos	■		■	■	
Sumarização	■			■	
Tesauros	■	■			■

Apesar dos vários comentários sobre a importância de se considerar características da língua na RI, até o presente momento, 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 freqüência das palavras, levam em consideração a co-ocorrência de palavras. Outros trabalham também com léxicos, bases de conhecimento e ontologias. 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 isso 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;

(ii) acredita-se que o custo-benefício de se 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. Isso 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 não possui morfologia tão rica quanto outras línguas. Características essas que deveriam ser consideradas não só quando se pensa em métodos lingüísticos, mas também no desempenho de métodos puramente estatísticos.

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, também disponíveis para o português, nossas pesquisas de PLN não foram ainda aplicadas em toda sua potencialidade na RI. Alguns exemplos de estudos/ferramentas de PLN para português que poderiam ser utilizados na RI são: análise morfossintática (Aires, 2000); Análise sintática (Bick, 2000; Martins *et al*, 2003); 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); resposta automática a perguntas (Lopes & Quaresma, 1999; Cunha, 1997); recuperação de informação geográfica (Padilha, 1997); co-referência (Rocha, 1999; Sant´Anna, 2000); sumarização automática (Pardo *et al*, 2003). Fontes mais completas de recursos e ferramentas de PLN disponíveis estão nos anais do evento Processamento Computacional do Português Escrito e Falado (Propor), e o site da Linguateca.[16]

Atualmente, existem várias iniciativas de código aberto para busca, o que permite que pesquisadores de PLN possam finalmente se aventurar sem maiores esforços na aplicação das técnicas e recursos por eles desenvolvidos na tarefa de busca. Alguns exemplos são: Lucene,[17] Nutch,[18] Carrot^2,[19] Objectssearch[20] e Egothor.[21] Lucene é uma biblioteca de código aberto para busca. Nutch é uma aplicação de código aberto para a busca na Web que utiliza Lucene, porém não é um site de busca, e pode ser utilizado tanto para a Web como para Intranets. Carrot^2 é um

[16] www.linguateca.pt
[17] http://jakarta.apache.org/lucene/
[18] http://www.nutch.org
[19] http://sourceforge.net/projects/carrot2/
[20] http://www.objectssearch.com/en/about.html
[21] http://www.egothor.org/

componente para meta-busca. Objectsearch é uma máquina de busca de código aberto *framework* para clusterização, que inclui além dos componentes para essa tarefa, um que utiliza Nutch, Lucene, Carrot^2, e outras bibliotecas de código aberto. Egothor também é uma máquina de busca de código aberto, que pode ser configurada como uma máquina *standalone*, como um meta-buscador, como um *HUB peer-to-peer* ou como uma biblioteca para outras aplicações que precisem de busca em textos.

II

Distinções mais sutis: para além do conteúdo

4. O problema do excesso de resultados irrelevantes

"The "why" of user search behavior is actually essential to satisfying the user's information need. After all, users don't sit down at their computer and say to themselves, "I think I'll do some searches."" Rose & Levinson (2004)

A Web se tornou um meio de informação e comunicação dos mais utilizados e mais importantes, devido a seu tamanho, taxa de crescimento e popularidade (Bar-Ilan & Peritz, 2004). Ferramentas ideais para lidar com todo esse volume de informação seriam ferramentas que levassem em conta em seu desenvolvimento todo o processo de busca de informação (*information seeking*), seus aspectos psicológicos, sociais e cognitivos. Por exemplo, motivações psicológicas independentes de contextos, efeito do contexto social no comportamento de busca, estágios no desenvolvimento de uma necessidade de busca, como solucionamos problemas, quais os efeitos da familiaridade com os tópicos da busca e da complexidade das tarefas a serem executadas. Alguns exemplos de emoções que podem influenciar em nossas buscas são: incerteza, otimismo, confusão, frustração, confidência e satisfação (Kalbach, 2003). Todos esses pontos vêm sendo investigados nas últimas décadas em estudos com usuários (Fidel *et al*, 2004) sobre interação humano-informação (*human-information behavior/interaction*), por pesquisadores que consideram a busca um processo cognitivo em que elaboramos uma idéia, podemos mudar de idéia, reformular nossas consultas, aprender enquanto procuramos, etc.

Entretanto, as máquinas de busca, principais e mais utilizadas ferramentas que temos livremente à disposição na Web para encontrar a informação desejada dado o grande volume disponível, não levam em consideração o processo de busca como um processo cognitivo nem as suas implicações. Também não consideram a complexidade das tarefas reais nas quais aplicamos as buscas (veja, por exemplo, Fidel *et al* (2004) sobre os aspectos envolvidos na busca colaborativa e Hirsh & Dinkelacker (2004) sobre o uso da busca para produzir informação). Não consideram nem mesmo diferenças mínimas nos processos de busca como, por exemplo, a diferença no processo de busca de uma criança e de um adulto (veja Bilal & Kirby (2001) para mais detalhes sobre esse assunto). Ao contrário, são os usuários que se

adaptam às máquinas de busca. Veja, por exemplo, os inúmeros sites com instruções sobre como obter melhores resultados de buscas[22].

De acordo com nossa revisão bibliográfica e análise das funcionalidades das principais máquinas de busca online até a entrega desta tese, além de não possibilitarem diferentes alternativas para diferentes comportamentos de busca, os quatro principais problemas das máquinas de busca são:

(i) Não ter toda a informação disponível na Web indexada. As máquinas de busca cobrem apenas parcialmente a Web (Bharat & Broder (1998), Lawrence & Giles(1998,1999)). Isso porque a Web tem crescido muito mais rápido do que qualquer método de indexação possa indexar. Não se consegue indexar páginas que são geradas dinamicamente, alguns sites têm acesso restrito apenas a certos usuários. Além disso, a sobreposição (*overlapping*) entre as diferentes máquinas de busca é pequena (Bharat & Broder, 1998). Veja, por exemplo, que Aires & Santos (2002) verificaram em 2002 que o Google tinha um número de páginas em português indexado cerca de 5 vezes menor do que o Alltheweb.

(ii) Não dar garantias sobre a qualidade da informação. Pois, páginas podem, por exemplo, ter sido retiradas da Web, conter conteúdo repetido, serem spam, serem atualizadas constantemente e não terem sido atualizadas no índice (Ntoulas *et al*, 2004).

(iii) Não retornar todas as páginas disponíveis na Web. Isso porque a consulta pode ter sido mal formulada, por exemplo, pela utilização de palavras muito comuns ou pouco descritivas, ou não utilização de todos os possíveis termos para uma determinada tarefa (sinônimos) ou porque as páginas relevantes estão em uma língua diferente da utilizada na consulta.

(iv) Não auxiliar o usuário a lidar com um volume muito grande de informação.

Dentre os quatro problemas acima, o terceiro e o quarto são os que têm soluções mais dependentes da língua em que os documentos devem ser recuperados.

[22] http://www.brightplanet.com/deepcontent/tutorials/search/index.asp;
http://Websearch.about.com/od/internetresearch/a/searchmistakes.htm?nl=1;
http://searchenginewatch.com/searchday/article.php/2159561

Como citado na Introdução, em novembro de 2002, Aires & Santos (2002) estimaram o tamanho da Web em português em 20.807.956 páginas no Alltheweb, 7.152.022 páginas no Altavista e 4.260.000 páginas no Google. Em junho de 2005, o mesmo experimento foi replicado e encontramos números de páginas bem superiores, que foram, respectivamente 149.000.000, 167.000.000 e 19.100.000 páginas. Um indicativo de que um número cada vez maior de páginas em português está disponível. Com a Web em português cada vez maior, cresce o número de estudos em aplicações de PLN para RI para português e RI entre línguas envolvendo o português (alguns exemplos são Silva & Oliveira, 2003; Gomes & Silva, 2005; Cardoso, Silva & Costa, 2004; Martins & Silva, 2004; Gonzalez et al, 2005; González & Lima, 2001; Pizzato & Lima, 2003; Orengo & Huyck 2002; Orengo & Huyck 2001).

O quarto problema pode ter duas causas: a pesquisa realmente trata de assunto muito abordado na Web para o qual muitos documentos relevantes são encontrados, ou existem muitos documentos sobre o assunto pesquisado, mas que não têm o foco que o usuário deseja para uma consulta em um dado instante.

Nos primeiros anos da Web, quando ainda era restrita a poucos grupos, tais como pesquisadores e estudantes, e a bases de dados específicas e técnicas como a Medline, era lógico que se assumisse que o único objetivo de um usuário era obter informações sobre um determinado tópico de pesquisa. Porém, atualmente, qualquer análise simples de *logs* de máquinas de busca pode mostrar que os objetivos de busca são muitos e bastante diversificados (Rose & Levinson, 2004). Apesar disso e dos diversos levantamentos dos assuntos procurados pelos usuários e seus comportamentos de busca (Jansen & Spink, 2005; Jansen *et al*, 2005; Jansen *et al*, 2000; Spink *et al*, 2000), poucos são os trabalhos que tentam levantar os porquês por trás de uma busca. Uma das poucas exceções é a taxonomia de tipos de busca de Broder (2002): navegacional (*navigational*), informacional (*informational*), transacional (*transactional*). O primeiro tipo se refere a buscas cujo objetivo é encontrar um site particular de onde se pode começar a navegar; o segundo se refere a encontrar informações sobre um dado tópico, que pode estar presente em um ou mais

resultados; e o último está relacionado a executar alguma atividade que é mediada por um determinado site, como, por exemplo, compras online.

Dois estudos mais recentes na mesma linha são os estudos de Aires & Aluísio (2003), em que as autoras, através de um experimento com alunos, professores, e funcionários de uma universidade brasileira, analisaram como os usuários expressam seus objetivos em suas consultas; e o estudo baseado na análise de *logs* de Rose & Levinson (2004), que diz que consultas podem ser classificadas quanto ao objetivo de busca como: navegacional, informacional ou recurso (*resource*). No primeiro caso, o objetivo é visitar a página de uma determinada instituição ou organização. No segundo, o foco é obter informações sobre um dado tópico, informações que podem ser classificadas segundo cinco necessidades: responder a uma pergunta (*direct*), obter conselhos (*advice*), aprender mais sobre um tópico (*undirected*), localizar algo no mundo real (*locate*) ou obter uma lista para pesquisas futuras (*list*). O último tipo se refere à necessidade de obter algo que não é informação, por exemplo, fazer *download* de um software.

Máquinas de busca, recentemente, anunciaram várias iniciativas para personalizar os resultados de busca, tentando aproximar os resultados do foco desejado pelo usuário, por exemplo:

(i) armazenando suas consultas e resultados de buscas como uma espécie de *bookmark* que proporcionam ao usuário uma pequena porção organizada da Web, ou até mesmo gravando parte do comportamento do usuário durante suas buscas (o que é aberto, impresso, etc.);

(ii) permitindo que o usuário especifique assuntos ou áreas que são particularmente interessantes para ele e, também, o seu grau de interesse sobre as mesmas;

(iii) permitindo que o usuário utilize *tabs*/atalhos para fazer buscas verticais, como, por exemplo, busca local, por notícias ou por pessoas famosas.

Essas iniciativas, ainda que interessantes, encontram os seguintes problemas: questões relacionadas à privacidade; como manter *profiles* por longos períodos e para diferentes tarefas; o foco de interesse do usuário pode mudar; suas buscas podem ser

muito específicas, por exemplo, relacionadas a uma tarefa particular que o usuário executa em seu trabalho.

O objetivo deste trabalho de doutorado foi pesquisar uma maneira de minimizar consideravelmente um dos principais problemas 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.

Partimos do pressuposto de que, para que um documento seja relevante, não basta que ele trate do assunto procurado, é necessário ainda que dê o enfoque desejado pelo usuário. O enfoque pode ser determinado por características como, por exemplo, formalidade, objetividade e o fato de o texto ser detalhado, tratando apenas de um assunto e não de vários.

A solução explorada neste trabalho para auxiliar o usuário a interpretar qual o enfoque sobre o assunto procurado é dado por um determinado texto foi classificá-los em gêneros, tipos de textos, necessidades de busca e necessidades personalizadas. Em nossa revisão da literatura, encontramos trabalhos que defendessem taxonomias de necessidades de busca, como o de Broder (2002) e o de Rose & Levinson (2004), que classificassem consultas segundo necessidades, como o de Kang & Kim (2003) que as classifica como buscando informações ou procurando *home pages*, porém, não encontramos trabalhos que classificassem textos como atendendo ou não a uma necessidade.

Para gerar os métodos de classificação, utilizamos marcadores de estilo/estilísticos, algoritmos de aprendizado de máquina e *corpora* compilados com textos em português.

O estilo é dado pela forma como o conteúdo de um texto é comunicado, pelas escolhas feitas por um autor devido a suas preferências pessoais, a forma como vê o leitor, ou a características que ele conhece e gosta sobre textos similares. Tais escolhas podem estar refletidas na forma como o texto foi organizado, em escolhas de vocabulário e sintáticas, ou na audiência para a qual o texto está voltado. Algumas

vezes, as variações são pequenas, mas o resultado dessas escolhas pode ser determinante para o sucesso da comunicação.

No próximo capítulo, definimos estilo e marcadores estilísticos, mostramos exemplos de tarefas que fazem uso de marcadores estilísticos, definimos gênero e resumimos trabalhos de classificação automática de textos, segundo gêneros que utilizam tais marcadores.

5. Estilo

" Texts are much more than what they are about." Karlgren (2004)

Ao nos expressarmos, obedecemos a diversas normas gramaticais e de uso da língua. Ainda assim, o uso da língua envolve uma quantidade significante de escolhas, por exemplo, a escolha de palavras e de construções sintáticas. Escritores e falantes podem selecionar uma variedade de palavras e outras características lingüísticas para passar suas mensagens para a audiência pretendida. As escolhas lingüísticas são feitas baseando-se tanto em características pessoais do escritor (por exemplo, dialeto) como em restrições contextuais (por exemplo, a mídia).

O estudo das escolhas lingüísticas é chamado de estilística (*stylistics*). Em sua definição mais geral, estilística é o estudo de qualquer uso situacional distintivo da língua, e das escolhas feitas por indivíduos e grupos sociais em seus usos da língua (Crystal, 1992, p. 371 *apud* Glover, 1996). O objetivo da estilística é identificar características estilisticamente significantes da língua (marcadores estilísticos ou de estilo), e as funções de certa forma por elas delimitadas (Crystal, 1992, p. 371 *apud* Glover, 1996).

Várias ênfases têm sido dadas ao conceito de estilo, em parte devido ao fato de que várias disciplinas que estudam estilo, como literatura, lingüística e filologia, dando exemplos apenas de disciplinas diretamente relacionadas ao estudo da língua, o estudam sob pontos de vista diferentes. Enkvist *et al* (1974) apresenta uma discussão detalhada sobre diversas definições de estilo.

Neste trabalho, adotamos o conceito de informação estilística como sendo a parte da informação lingüística que não está relacionada ao conteúdo, mas sim à forma como o conteúdo é transmitido/comunicado. Exemplificando, a oração "Por favor, passe o sal" contrasta estilisticamente com a oração "Por favor, dê-me 10 miligramas de cloreto de sódio". As duas orações têm probabilidades de aparecer em contextos relacionados, porém diferentes, respectivamente, à sala de jantar e a um

laboratório. Entretanto, elas não contrastam estilisticamente com a oração "Por favor, passe a pimenta" (Enkvist *et al*., 1974).

Algumas das informações estilísticas são características individuais nossas enquanto autores e servem como marcadores do estilo do autor, outras são descritas em manuais de redação, ensinadas em treinamento profissional (por exemplo, muitas vezes seguimos um certo padrão da empresa para escrever nossos relatórios) e servem como marcadores do estilo de um grupo.

O estilo seria, então, um fator de individualidade, como uma expressão distintiva de um autor, grupo, sociedade, ou uma combinação desses e pode ser definido como um conjunto de propriedades que permitem a especialistas falar sobre gêneros e a pessoas em geral a acessar o que é apropriado ou não em contextos específicos (Aires *et al*, 2005a).

5.1 Estilometria

Estilometria (*stylostatistics*/ *stylometry*/ *statistical stylistics/ computational stylistics*) é uma área crescente dentro da estilística (Glover, 1996), que lida com a análise quantitativa do estilo de escrita. O objetivo dessa área é encontrar informações a respeito de estilo através de características contáveis. Para ela, estilo é um conceito probabilístico, e tendências estilísticas poderiam ser reveladas a partir de exemplos de textos variados.

5.1.1 Aplicações da estilometria

A estilometria tem sido utilizada principalmente em três áreas: descrição de características estilísticas de períodos históricos específicos, incluindo a investigação de mudanças diacrônicas[23] na língua; identificação de características de estilos de escrita de um particular autor; e procura de conjuntos de marcadores estilísticos associados a diferentes gêneros. Para tais áreas, a estilometria está tanto relacionada a estudos teóricos de cunho histórico e analítico, como de cunho prático, para

[23] Estudos diacrônicos contrastam com sincrônicos, esses últimos tratando de análises em um período de tempo fixo, enquanto que os primeiros tratam de análises em vários períodos (décadas ou séculos, por exemplo), visando a estudar a evolução ou explicação de um fenômeno, por exemplo.

aplicações em sistemas computacionais. Exemplos de estudos de cunho teórico/analítico são a análise cronológica do trabalho de um determinado autor; a determinação da data de um determinado trabalho (*stylochronometry*) e de marcadores estilísticos de autores específicos para determinação de autoria (*authorship attribution* ou *fingerprinting*). Can & Patton (2004), por exemplo, estudam as mudanças de estilo de dois autores turcos, Çetin Altan e Yaşar Kemal.

Algumas das motivações por trás de casos de dúvida de autoria são:

(i) Os autores nem sempre tiveram os direitos autorais sobre suas produções, muitas vezes os direitos eram dos editores ou dos donos dos textos. Na Inglaterra, por exemplo, até o século XVIII, os direitos editoriais eram preservados, mas não os direitos dos autores, que muitas vezes tinham seus nomes deliberadamente não mencionados nos textos;

(ii) Nomes de autores famosos eram utilizados em trabalhos de terceiros, por exemplo, na Grécia antiga, para garantir que os trabalhos fossem lidos;

(iii) Alguns autores deliberadamente publicaram utilizando pseudônimos, por exemplo, para evitar críticas ou por estarem publicando em um gênero diferente do que habitualmente publicavam. Rudman (2002) e Stamatatos *et al* (1999, 2000) apresentam exemplos de estudos para determinação de autoria.

Outro exemplo de estudo teórico/analítico é descrito em Argamon *et al* (2003), onde se trata da diferença entre a escrita de mulheres e homens no *British National Corpus*[24] (BNC). Os autores apresentam conclusões como, por exemplo, mesmo em textos formais as mulheres utilizavam mais características descritas em outros trabalhos sobre estilo como pessoal/emotiva (*involved*) e homens mais características descritas como informativas/impessoais (informational).

Exemplos de aplicações práticas para estudos sobre autoria podem ser vistos na área de Lingüística Forense (*Forensic linguistics*) e em sistemas de escrita colaborativa. A lingüística forense utiliza técnicas lingüísticas para investigar crimes em que dados escritos são partes das evidências (Crystal, 1992, p. 142 *apud* Glove, 1996). Por exemplo, para distinguir confissões genuínas de confissões forjadas por

[24] http://www.natcorp.ox.ac.uk/

policiais; para determinar a autoria de cartas anônimas, testamentos e outros documentos legais (Glove, 1996). No caso da escrita colaborativa (*collaborative writing*) (Baljko & Hirst, 1999; Glover & Hirst, 1996), o objetivo é identificar os diferentes estilos dos autores para que se possa mesclá-los/uni-los (fazer um único estilo de vários estilos) de forma harmônica, uma vez que variação de estilo pode prejudicar o entendimento do texto e levar o leitor a conclusões como a de que o texto foi feito sem maiores cuidados (Glover & Hirst, 1996). Um sistema para auxílio à escrita colaborativa envolve duas fases: (i) descobrir as inconsistências estilísticas e (ii) apresentar os resultados mostrando os problemas encontrados e sugestões de mudanças para que os autores possam corrigir as inconsistências (Glover & Hirst, 1996). No Quadro 5, mostramos alguns dos marcadores de estilo para identificação de autoria mostrados por Glover & Hirst (1996) como marcadores passíveis de serem aplicados para a tarefa de escrita colaborativa[25].

Outro exemplo de aplicação prática de estilometria é em sistemas tutores para ensino de escrita (*inteligent tutoring systems for teaching writing*). Um exemplo de sistema tutor desse tipo é o STASEL (Stylistic Treatment at the sentence level) que ensina princípios de estilo sintático para estudantes de inglês (Payette & Hirst, 1992).

Aplicações práticas podem ser vistas também em sistemas de processamento de linguagem natural, como sistemas de geração de textos, e de tradução automática (DiMarco & Hirst, 1990; DiMarco, 1990). No caso dos sistemas de geração, poderiam produzir textos que dizem a mesma coisa, mas distintos estilisticamente para atender a diferentes audiências e contextos.

A importância do uso de estilo na tradução automática se deve ao fato de que parte significativa do significado de um texto é dada pelo estilo do autor, uma vez que escolhas diferentes de palavras e estruturas sintáticas cobrem diferentes nuances do significado e, portanto, deveriam ser trazidas para a tradução para que se preservasse a intenção do autor e o texto traduzido pudesse ser fiel ao texto original. O uso do estilo colabora também para que, mesmo preservando-se o estilo do autor, o texto gerado tenha características naturais de estilo para o público alvo. Por exemplo,

[25] Neste texto, as listas de marcadores de estilo utilizadas em trabalhos do inglês serão mantidas em inglês, pois nem todos os marcadores possuem equivalentes em português.

estudos comparativos de estilística mostram que a língua francesa naturalmente aborda conceitos mais teóricos enquanto que o inglês tem uma predileção sobre a descrição de pormenores concretos. Assim, e para manter o mesmo nível estilístico numa tradução de francês para inglês, um texto original abstrato e aparentemente formal deverá ser traduzido de uma forma mais concreta e direta em inglês (DiMarco & Hirst, 1990).

Quadro 5 - Marcadores de estilo para identificação de autoria que podem ser aplicados para a tarefa de escrita colaborativa (Glover & Hirst, 1996)

Texto cru
• Register of words used (formal, slang, technical, etc.)
• Frequent words (at least 3 per thousand)
• Sentence length (mean and standard deviation)
• Word length (mean and standard deviation)
Texto etiquetado morfossintaticamente
• Type / token ratio
• Distribution of word classes (parts of speech)
• Distribution of verb forms (tense, aspect, etc.)
• Frequency of word parallelism
• Distribution of word-class patterns (e.g., determiner + noun + verb)
• Distribution of nominal forms (e.g., gerunds)
• Richness of vocabulary
Texto etiquetado sintaticamente
• Frequency of clause types
• Distribution of direction of branching
• Frequency of syntactic parallelism
• Distribution of genitive forms (of and 's)
• Distribution of phrase structures
• Frequency of imperative, interrogative, and declarative sentences
• Frequency of topicalization
• Ratio of main to subordinate clauses
• Distribution of case frames
• Frequency of passive voice
Texto etiquetado semanticamente
• Frequency of negation
• Frequency of deixis
• Frequency of hedges and markers of uncertainty
• Frequency of semantic parallelism
• Degree of alternative word use (preference for synonyms)

Outra aplicação prática de estilometria é a classificação de textos segundo gêneros, que será descrita em mais detalhes na Seção 5.2.

Nas diversas aplicações de estilometria mencionadas nesta Seção, nos deparamos com problemas de classificação supervisionada, dado que temos um conjunto de objetos pré-classificados (segundo autores, gêneros, períodos de tempo,

etc.) e tentamos categorizar novos objetos (textos) como pertencentes a um desses conjuntos. As aplicações mais modernas de estilometria utilizam para esses casos técnicas de aprendizado de máquina e os marcadores de estilo como *features* para o treinamento dos classificadores/categorizadores de texto (Whitelaw & Argamon, 2004). Tradicionalmente, os modelos gerados para classificação nesses casos são baseados em conjuntos de marcadores de estilo lexicais, sintáticos e independentes de conteúdo, selecionados manualmente.

5.1.2 Marcadores de estilo

Em geral, através da freqüência dos marcadores de estilo em um texto, podemos tecer conclusões quanto a características como formalidade, elegância, complexidade sintática e complexidade lexical de um texto. Tais características são resultados das escolhas do autor de acordo com o gênero, a mídia, o propósito do texto, os níveis de educação e social e a personalidade do autor e da audiência (Whitelaw & Argamon, 2004) e, por sua vez, caracterizam o estilo de um texto, como, por exemplo, em estilo científico, jornalístico, de comunicação diária e estilo literário (Michos *et al*., 1996). Alguns exemplos de marcadores estilísticos são: (i) marcadores relacionados a palavras, como expressões idiomáticas, expressões sofisticadas, terminologia científica, palavras formais e abreviaturas; (ii) marcadores sintáticos, como o número de palavras por frase, número de conjunções por frase, número de sentenças por parágrafo, proporção de verbos versus substantivos, porcentagem de verbos na terceira pessoa, porcentagem de orações subordinadas e proporção de adjetivos versus substantivos.

Textos formais seriam, por exemplo, caracterizados pelo grande uso de palavras formais e expressões sofisticadas, pela pouca freqüência de abreviações e expressões idiomáticas, por um número alto de palavras por frase, número pequeno de frases por parágrafo, um número alto de conjunções por frases, uma porcentagem alta de verbos na terceira pessoa e voz passiva predominante (Michos *et al*., 1996). Por sua vez, a elegância seria, por exemplo, caracterizada por muitas expressões idiomáticas, palavras de uso poético, alta porcentagem de pares de adjetivos e substantivos, de pares de advérbios e verbos e predominância da voz ativa (Michos *et al*., 1996). Já a complexidade sintática se daria, por exemplo, através de um número

alto de palavras por frase, muitas frases por parágrafo, muitas conjunções por frase e alto número de orações subordinadas (Michos *et al*., 1996). Textos lexicalmente complexos, por sua vez, seriam caracterizados por várias expressões sofisticadas, bastante terminologia científica, muitas palavras formais, muitas abreviações e palavras de uso poético e algumas expressões idiomáticas (Michos *et al*., 1996).

Marcadores de estilo são, portanto, traços lingüísticos, independentes de tópico/assunto que servem como medidas objetivas para análise dos objetivos estilísticos do autor em um dado texto. São aqueles itens lingüísticos que apenas aparecem em determinados contextos, ou que são muito ou pouco freqüentes neles, ou seja, são elementos lingüísticos contextualmente limitados. Tais marcadores podem se manifestar em termos de metro[26] (por exemplo, dísticos[27] heróicos[28]), de tempo (por exemplo, estilo isabelino[29]), de lugar (por exemplo, humor ianque[30]), de linguagem, de dialeto, de escritor (por exemplo, estilo byroniano[31]) ou de obra literária (por exemplo, eufuísmo[32]), de escola literária (por exemplo, estilo romântico), de gênero (por exemplo, linguagem poética e jornalística), de situação social (por exemplo, o sargento dirigindo-se ao recruta) e assim por diante (Enkvist, 1974).

5.1.3 A escolha de marcadores de estilo

A escolha de marcadores de estilo não é baseada em teorias lingüísticas e é tradicionalmente realizada de forma manual (não automática), com exceção do caso de trabalhos como o Whitelaw & Argamon (2004) que utilizam como marcadores as palavras mais freqüentes, isto é, palavras funcionais (*function words*)[33].

Segundo Enkvist (1974), a escolha manual dos marcadores pode se dar de duas formas: (i) através do método clássico dos críticos literários de confiar em sua

[26] Forma rítmica de obra poética.
[27] Grupo de dois versos.
[28] Sátira em prosa como, por exemplo, as feitas pelos escritores ingleses Alexander Pope e Jonathan Swift
[29] Produção realizada durante o reinado de Elisabete I ou Isabel I.
[30] De New England (E.U.A.), região cultural e lingüística constituída pelos estados de Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island e Vermont.
[31] Do poeta inglês Lord Byron, que influenciou fortemente a geração seguinte na Inglaterra.
[32] Estilo literário afetado, amaneirado, que se usou na Inglaterra ao tempo da rainha Isabel I.
[33] Também traduzido como palavras gramaticais.

experiência ou "sensibilidade ao estilo" para definir quais expressões são comuns e quais não o são, que poderia ser também feita com o auxílio do computador para avaliar a freqüência de tais expressões; (ii) utilizando-se de um grupo de informantes.

No primeiro caso, o estudioso deve ter cuidado para não misturar as considerações tecidas a respeito da freqüência das expressões com outras considerações e analisar prévia e cuidadosamente o contexto que será estudado. No caso de a tarefa ser feita com o auxílio do computador, pode-se ainda analisar os padrões de ocorrência dos marcadores estilísticos inicialmente selecionados, utilizando métodos estatísticos, como análise de componentes principais (*principal component analysis*) (Smith, 2002); análise discriminante (*discriminant analysis*) (McLachlan, 2004) e regressão lógica (*logistic regression*) (Hosmer & Lemeshow, 2000). No segundo caso, é possível, por exemplo, dar a um grupo de informantes uma passagem de texto como um estímulo ou um esquema, e a seguir perguntar-lhes qual item (ou itens lingüísticos) esperam que ocorra(m) em seguida ou próximo ao texto. Inversamente, pode-se pedir aos informantes que definam contextos nos quais uma dada expressão tem probabilidade de aparecer. O número de palpites certos, ou de expectativas realizadas, dá uma medida aproximada da previsibilidade relativa dos itens conjeturados. Esse processo requer uma distinção inicial entre escolhas gramaticais não-estilísticas e estilísticas, tanto quanto um certo controle de fatores contextuais antes da realização de uma avaliação de previsibilidades estilísticas. Na prática, os informantes devem ser cuidadosamente selecionados em relação à educação, experiência e capacidade lingüística.

Em nossa revisão da literatura, não encontramos trabalhos que utilizassem o segundo método descrito acima para selecionar marcadores de estilo. Em todos os trabalhos por nós revistos, a escolha de marcadores de estilo era feita ou através de intuição lingüística ou do uso de marcadores de estilo definidos como úteis por trabalhos de autores que haviam feito uso de sua intuição lingüística.

Um exemplo do uso de intuição lingüística para seleção inicial de marcadores de estilo é o trabalho de Biber (1988), que faz uso dela para reunir os marcadores que encontrou citados na revisão bibliográfica de mais de 200 estudos. Biber estudou a variação de textos em inglês, usando diversas variáveis, e descobriu que diferentes

registros se diferenciavam sistematicamente ao longo de cinco dimensões, relacionadas a considerações funcionais como interatividade (*interactiveness*), envolvimento (*involvement*), propósito (*purpose*) e circunstâncias de produção; todas essas considerações com seus marcadores correlatos na estrutura lingüística. Veja na Tabela 3 (Biber, 1993) as cinco dimensões, seus marcadores estilísticos e os registros característicos. Posteriormente, Biber (1995), através de estudos comparativos para outras línguas, identificou que tais marcadores de estilo seriam independentes de língua. O Quadro 6 apresenta os marcadores de estilo levantados por Biber. Biber foi o primeiro a estudar a interação (positiva e negativa) entre um grande conjunto de fatores (marcadores estilísticos) que até então tinham apenas sido mencionados individualmente por outros autores; seu trabalho descritivo e empírico é pioneiro.

Tabela 3 – Dimensões e seus marcadores estilísticos (Biber, 1993, p. 231-232)

Funções	**Características lingüísticas**	**Registros característicos**
Dimensão 1 – Produção informacional versus produção com interação/com envolvimento		
Monólogo, produção cuidadosa, informativa, sem envolvimento	*Nouns, adjectives, prepositional phrases, long words*	Exposição informativa, e.g., documentos oficiais e prosa acadêmica
Interativa, foco interpessoal, envolvida, postura pessoal, produação online	1st and 2nd person pronouns, questions, reductions, stance verbs, hedges, emphatics, adverbial subordination	Conversação, e.g., cartas pessoais e conversas públicas
Dimensão 2 – Produções narrativas versus não-narrativas		
Narrativa	*Past tense, perfect aspect, 3rd person pronouns, speech act (public) verbs*	Ficção
Não narrativa	*Present tense, attributive adjectives*	Exposição, rádio, cartas profissionais, conversas telefônicas
Dimensão 3 – Referências explícitas versus referências dependentes de contexto		
Referência independente de contexto, elaborada	*WH relative clauses, pied-piping constructions, phrasal coordination*	Documentos oficiais, cartas profissionais, exposição
Referência dependente de contexto, produção online	*Time and place adverbials*	Rádio, conversação, ficção, cartas pessoais
Dimensão 4 – Expressão explícita de persuasão versus não-explícita		
Argumentação e persuasão explícitas	*Modals (prediction, necessity, possibility), suasive verbs, conditional subordination*	Cartas profissionais, editoriais
Argumentação não explícita	——	Rádio, críticas literárias
Dimensão 5 – Informação abstrata versus não-abstrata		
Estilo abstrato	*Agentless passives, by passives, passive dependent clauses*	Prosa técnica, outras prosas acadêmicas, documentos oficiais
Não abstrato	——	Conversação, ficção, cartas pessoais, discursos, conversas públicas, rádio

Quadro 6 - 67 Marcadores de estilo levantados por Biber para o inglês (Biber, 1995, p. 95-96)

Tense and aspect markers
1. *Past tense*
2. *Perfect aspect*
3. *Present tense*

Place and time adverbials
4. *Place adverbials*
5. *Time adverbials*

Pronouns and pro-verbs
6. *First-person pronouns*
7. *Second- person pronouns*
8. *Third- person personal pronouns (excluding it)*
9. *Pronoun it*
10. *Demonstrative pronouns (that, this, these, those as pronouns)*
11. *Indefinite pronouns (e.g., anybody, nothing, someone)*
12. *Pro-verb do*

Questions
13. *Direct WH questions*

Nominal forms
14. *Nominalizations (ending in –tion, -ment, -ness, -ity)*
15. *Gerunds (participial forms functioning as nouns)*
16. *Total other nous*

Passives
17. *Agentless passives*
18. *By-passives*

Stative forms
19. *Be as main verb*
20. *Existential there*

Subordination features
21. *That verb complements (e.g. I said that he went)*
22. *That adjective complements (e.g. I'am glad that you like it)*
23. *WH-clauses (e.g., I belived what he told me)*
24. *Infinitives*
25. *Present participal adverbial clauses (Stuffing his mouth with cookies, Joe run out the door)*
26. *Past participal adverbial clauses (Built in a single week, the house would stand for fifty years)*
27. *Past participal postnominal (reduced relative) clauses (The solution produced by this process…)*
28. *Present participal postnominal (reduced relative) clauses (The event causing this decline was…)*
29. *That relative clauses on subject position (The dog that bit me)*
30. *That relative clauses on object position (The dog that I saw)*
31. *WH relatives on subject position (The man who likes popcorn)*
32. *WH relatives on object position (The man who Sally likes)*
33. *Pied-piping relative clauses (The manner in which he was told)*
34. *Sentence relatives (Bob likes fried mangoes, which is the most disgusting thing I've ever heard of)*
35. *Causative adverbial subordinator (because)*
36. *Concessive adverbial subordinators (although, though)*
37. *Conditional adverbial subordinators (if, unless)*
38. *Other adverbial subordinators (since, while, whereas)*

Prepositional phrases, adjectives, and adverbs
39. *Total prepositional phrases*
40. *Attributive adjectives (The big horse)*
41. *Predicative adjectives (The horse is big)*
42. *Total adverbs*

Lexical specificity
43. *Type-token ratio*
44. *Mean word length*

Lexical classes
45. *Conjuncts (consequently, furthermore, however)*
46. *Downtoners (barely, nearly, slightly)*
47. *Hedges (at about, something like, almost)*
48. *Amplifiers (absolutely, extremely, perfectly)*
49. *Emphatics (a lot, for sure, really)*
50. *Discourse particles (sentence-initial well, now, anyway)*
51. *Demonstrative*

Modals
52. *Possibility modals (can, may, might, could)*
53. *Necessity modals (ought, should, must)*
54. *Predictive modals (will, would, shall)*

Specialized verb classes
55. *Public verbs (assert, declare, mention)*
56. *Privative verbs (assume, believe, doubt, know)*
57. *Suasive verbs (command, insist, propose)*
58. *Seem and appear*

Reduced forms and dispreferred structures
59. *Contractions*
60. *Subordinator that deletion (I think he went..)*
61. *Stranded prepositions (The candidate that I was thinking of...)*
62. *Split infinitives (He wants to convincingly prove that...)*
63. *Split auxiliaries (They were apparently shown to…)*

Co-ordination
64. *Phrasal co-ordination (noun-noun; adj; and adj; verb and verb; adv and adv)*
65. *Independent clause co-ordination (clause initial and)*

Negation
66. *Synthetic negation (No answer is good enough to Jones)*
67. *Analytici negation (That's not likely)*

5.2 Classificação de textos em gêneros

Na comunicação escrita, encontramos uma grande variação no estilo dos textos, por exemplo, uma carta, um relatório técnico, um *curriculum* têm formas diferentes, que são de certa forma responsáveis pelas expectativas que teremos sobre o documento. Essas formas chamam nossa atenção de modos diferentes para o conteúdo do texto. Marcadores de estilo podem ser utilizados para distinguir diferentes gêneros (Biber, 1988), como jornalístico e literário, ou até mesmo para fazer distinções dentro desses gêneros, por exemplo, prosa versus drama.

O termo gênero é um conceito variável, uma vez que para uma dada área/assunto, por exemplo, música, filmes, literatura, não existe um conjunto fixo de categorias de gêneros. Identificar uma taxonomia de gênero é um processo subjetivo, no qual as pessoas podem discordar sobre o que constitui um gênero, ou dos critérios sobre a inclusão de um texto em uma determinada categoria. Pode-se dizer que o gênero reflete o estilo do texto, dando informações sobre que tipo de texto é, mas não fornece informações sobre o que é o texto, uma vez que gênero é ortogonal ao tópico dos textos. Textos sobre o mesmo assunto podem ser de diferentes gêneros e textos do mesmo gênero podem tratar de diferentes assuntos.

Em nossa revisão, encontramos sete trabalhos de classificação automática de textos[34] em gêneros que utilizam marcadores estilísticos: seis para o inglês (Kessler *et al*, 1997; Argamon *et al*, 1998; Karlgren, 2000; Stamatatos *et al*, 2000a; Dewdney *et al*, 2001; Finn *et al*, 2002) e um para o grego (Stamatatos *et al*, 2000b). Existem diversos trabalhos sobre classificação de textos, porém aqui tratamos apenas da classificação segundo gêneros. Nessa contagem nos restringimos ainda a trabalhos que: (i) a classificação fosse feita de forma automática; (ii) o conjunto de *features* incluísse marcadores de estilo; (iii) utilizassem um conjunto de classes previamente conhecido; (iv) chamassem as classes/categorias tratadas por eles de gêneros[35]; (v) os gêneros tratados não fossem apenas relacionados ao domínio, como, por exemplo,

[34] Sebastiani (2002) apresenta um bom panorama sobre classificação automática de textos, utilizando algoritmos de aprendizado de máquina.
[35] Ainda que não concordássemos que as categorias eram gêneros. Por exemplo, Finn *et al* (2002) trata críticas positivas e negativas como gêneros, porém, em nossa opinião essas categorias refletem apenas a atitude do autor perante o conteúdo.

matemática, medicina e direito. Por isso, não mencionamos trabalhos como o de Rauber & Müller-Kögler (2001) em que marcadores de estilo são utilizados para organizar livros em um sistema de biblioteca digital (Somlib) de acordo com suas semelhanças estruturais e sintáticas, utilizando mapas auto-organizáveis (*self-organizing map*) (SOM)[36] para formar clusters dos documentos e representá-los com diferentes cores, ou seja, os gêneros não são previamente determinados. Nem trabalhos como o de Roussinov *et al* (2001), em que 5 grupos de gêneros da Web são propostos, porém, não são realizados experimentos de classificação automática.

Nas próximas subseções, resumimos os sete trabalhos encontrados em nossa revisão. A diferença no nível de detalhes fornecido é devida à grande variação da quantidade de informações fornecidas nos artigos nos quais tais trabalhos são descritos. Por exemplo, a maioria dos autores apenas menciona alguns exemplos de marcadores de estilo utilizados, mas não fornecem a lista completa; alguns não fornecem sequer o número total de marcadores utilizados.

5.2.1 O trabalho de Kessler *et al* (1997)

Kessler *et al* (1997) reportam experimentos com três esquemas de classificação, utilizando 499 textos dos 802 textos do *corpus* Brown[37]: (i) de acordo com o nível intelectual da audiência (popular, médio, médio-superior e alto); (ii) se um texto é uma narrativa ou não, ou seja, se primariamente relata uma seqüência de eventos; (iii) reportagem, editorial, escrita científica ou técnica, textos sobre lei e administração do governo, não-ficção e ficção. Dos 499 textos, 402 são utilizados para treinamento e 97 para teste.

Em seus experimentos utilizam 55 *features*, tais como: palavras utilizadas para expressar data; afixos latinos; número de interrogações e número de palavras hifenizadas. Com esse conjunto de *features* são utilizados três métodos diferentes para gerar os modelos de classificação: regressão logística, redes neurais perceptron de duas camadas e de múltiplas camadas (Shepherd, 1997). A melhor taxa de acerto é de 54% para o primeiro esquema e de 86% para o segundo esquema utilizando uma rede

[36] http://koti.mbnet.fi/~phodju/nenet/SelfOrganizingMap/General.html

[37] http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

neural perceptron com múltiplas camadas; para o terceiro esquema, é de 75% utilizando uma rede perceptron com duas camadas.

5.2.2 O trabalho de Argamon *et al* (1998)

Argamon *et al* (1998) utilizam em seus experimentos um *corpus* com 800 textos, sendo 200 de cada uma das seguintes fontes: (i) notícias do jornal NY Times de janeiro de 1998; (ii) editoriais do jornal NY Times de janeiro de 1998, (iii) notícias do NY Daily News de janeiro de 1998; (iv) artigos da revista Newsweek sobre questões domésticas dos Estados Unidos, de julho 1997 a janeiro 1998. Nos experimentos são utilizados dois tipos de *features*: as freqüências de 500 palavras funcionais (tais como "*and*", "*about*" e "*the*") e 685 trigramas morfossintáticos (por exemplo: pronome pessoal + verbo no presente + verbo no gerúndio). Para selecionar os trigramas que foram utilizados como *features*, os textos foram etiquetados com o etiquetador de Brill (1994) utilizando o *tagset* do *corpus* Brown. Resultaram 100 mil trigramas, dos quais foram selecionados os que apareciam em, pelo menos, 25% dos textos do *corpus* e, no máximo, em 75% dos textos do *corpus*, ou seja, apenas 685.

Em seus experimentos são utilizados o algoritmo ripper (um tipo de algoritmo que utiliza árvores de decisão) (Cohen 1995, 1996) e validação cruzada estratificada com cinco partes (*five-fold cross-validation*). No artigo são apresentadas apenas as taxas de precisão considerando pares de gêneros. Os resultados são mostrados na Tabela 4.

Tabela 4 – Taxas de acerto apresentadas por Argamon *et al* (1998)

	Palavras funcionais	Trigramas	Ambos
Notícias x editoriais do Times	78%	69,3%	79,5%
Notícias do Times x Notícias do NY Daily News[38]	82,3%	63,1%	84,3%
Notícias do Times x artigos da Newsweek	80,5%	79,3%	83,8%
Editoriais do Times x artigos da Newsweek	61,3%	70%	68,5%
NY Daily News x editoriais do Times	77,6%	67,3%	78,1%
NY Daily News x artigos da Newsweek	78,5%	79,6%	80,6%

[38] Nesse caso em particular, a distinção não é entre gêneros, mas entre estilos editoriais de diferentes jornais.

5.2.3 O trabalho de Karlgren (2000)

Karlgren (2000) apresenta experimentos de classificação realizados com dois *corpora*: (i) o *corpus* Brown (Karlgren 2000, Capítulo 7); (ii) um *corpus* com 1358 textos extraídos da Web (Karlgren 2000, Capítulos 15 e 16). Seus experimentos são também descritos em Karlgren & Cutting (1994), Karlgren & Straszheim (1997), Dewe *et al*. (1998), Bretan *et al*. (1998) e Karlgren *et al*. (1998).

Com o *corpus* Brown apresenta resultados para três esquemas de classificação: (i) informativos (*informative*) e de ficção (*imaginative*); (ii) textos jornalísticos (*press*), miscelânea (*misc*), não-ficção (*non-fiction*) e ficção (*fiction*); e (iii) reportagens (*press-reportage*), editoriais (*press-editorial*), artigos (*press-reviews*), religião (*religion*), atividades e hobies (*skills and hobbies*), fábulas (*popular lore*), literatura clássica (*belles lettres, etc*.), documentos do governo e miscelânea (*gov. doc. & misc.*), textos científicos (*learned*), ficção (*general fiction*), mistério (*mystery*), ficção científica (*science fiction*), aventura e faroeste (*adv. & western*), romance (*romance*), e humor (*humor*). Utiliza 20 *features* baseadas no trabalho de Biber (1988), mostradas no Quadro 7.

Quadro 7 – Marcadores de estilo utilizados por Karlgren (2000, Capítulo 7, p. 65) em seus experimentos com o Brown *corpus*

1. Adverb count	11.Present participle count
2. Character count	12.Sentence count
3. Long word count (> 6 chars)	13.Type / token ratio
4. Preposition count	14.“I” count
5. Second person pronoun count	15.Character per word average
6. “Therefore” count	16.“It” count
7. Words per sentence average	17.Noun count
8. Chars / sentence average	18.Present verb count
9. First person pronoun count	19.“That” count
10.“Me” count	20.“Which” count

Para treinamento utiliza análise discriminante; as taxas de acerto para 2, 4 e 15 classes são, respectivamente, 95,6%, 73,2% e 51,6%.

O *corpus* da Web foi formado por: (i) os 10 primeiros resultados de máquinas de busca para os tópicos 251-300 do TREC, gerando um total de 386 documentos; (ii) os resultados para 60 consultas da lista de últimas consultas realizadas na máquina de busca Magellan[39], obtendo 478 documentos; (iii) 494 documentos do histórico

[39] Magellan era uma máquina de busca que foi comprada pela Excite.

(*history files*) de colegas. Os textos do *corpus* foram então classificados em um dos 11 gêneros mostrados no Quadro 8.

Quadro 8 – 11 Gêneros considerados por Karlgren (2000, Capítulo 15, p. 116)

Textos informais, ou privados, páginas pessoais
Textos públicos, comerciais, páginas para o público em geral
Páginas interativas e formulários
Texto jornalístico: notícias, reportagens, editoriais, críticas, *popular reporting*, *e-zines*.
Relatórios, textos legais, material público; textos formais.
Outros textos
FAQs
Coleções de links
Listas e tabelas
Correspondência assíncrona com muitos participantes: contributições para discusões, requerimentos, comentários; material de *Usenet News*.
Mensagens de erro

Para selecionar os 11 gêneros mostrados no Quadro 8, Karlgreen solicitou por *e-mail* a 648 estudantes e pesquisadores da Universidade de Estocolmo e do Royal Institute of Technology que dissessem quais gêneros eles encontravam na Web. 67 pessoas responderam ao *e-mail* e da lista de sugestões foram selecionados os gêneros que pudessem ser distinguidos com marcadores simples de estilo. A lista de gêneros foi então enviada por *e-mail* para o mesmo grupo de 648 pessoas, 102 delas enviaram respostas sobre a paleta de gêneros. Alguns dos comentários foram: (i) não entenderam o que eram "FAQ", "Listas e tabelas", "Mensagens de erro"; (ii) sugeriram a inclusão de páginas de *download* e FTP; (iii) sugeriram gêneros mais baseados em conteúdo; (iv) disseram que não conseguem se imaginar buscando por "mensagens de erro" ou "páginas interativas". Porém, todos os comentários foram ignorados e a lista foi mantida.

Karlgren não mostra a lista de marcadores utilizados nos experimentos com esse *corpus*; diz apenas que é baseada em marcadores utilizados em experimentos anteriores e que 40 deles aparecem nas 12 regras geradas com o algoritmo C4.5 (Quinlan, 1992). Também não apresenta a taxa de acerto. Diz apenas que pediu a 12 pessoas (6 homens e 6 mulheres, entre 25 e 30 anos, razoavelmente experientes em buscas) que fizessem duas buscas: uma utilizando Easify[40] (uma ferramenta de busca que apresenta *clusters* dos resultados e mostra os gêneros dos mesmos) e a outra a máquina de busca Altavista e que a maioria dos usuários utilizou a interface como

[40] http://www.sics.se/humle/projects/DropJaw/

deveria, procurando nos gêneros em que as consultas deveriam acontecer. As duas consultas utilizadas são: (i) Encontrar um álbum ou crítica a concerto do Oasis; e (ii) Encontrar uma lista de hotéis em Malta.

5.2.4 O trabalho de Stamatatos *et al* (2000a)

Stamatatos *et al* (2000a) utilizam o *corpus* Wall Street Journal (WSJ)[41] do ano de 1989. Apesar desse *corpus* não ser categorizado quanto a gêneros, ele possui etiquetas que podem ser utilizadas para determinar o gênero a que corresponde cada texto. Tais etiquetas foram utilizadas pelos autores para que os textos fossem classificados segundo os 4 gêneros: editoriais, cartas ao editor, reportagem e notícias curtas (*spot news*).

O conjunto de *features* utilizado consiste nas 75 palavras mais freqüentes do *corpus* BNC e 8 *features* relacionadas à freqüência de ocorrência dos símbolos de pontuação: ponto, vírgula, ponto e vírgula, aspas, parênteses, interrogação e hífen. Em seus experimentos os autores utilizaram de 5 a 75 das palavras mais freqüentes (variando de 5 em 5) e verificaram que não havia ganho após as 30 primeiras palavras. A taxa de acerto utilizando análise discriminante, as 8 *features* de pontuação e entre 15 e 35 *features* relacionadas à freqüência das palavras, fica em torno de 97%. No Quadro 9 mostramos as 50 palavras mais freqüentes do BNC *corpus*.

Quadro 9 - As 50 palavras mais freqüentes do BNC *corpus* (Stamatatos *et al*, 2000a, p. 810)

1. the	*11. with*	*21. are*	*31. or*	*41. her*
2. of	*12. he*	*22. not*	*32. an*	*42. n't*
3. and	*13. be*	*23. his*	*33. were*	*43. there*
4. a	*14. on*	*24. this*	*34. we*	*44. can*
5. in	*15. i*	*25. from*	*35. their*	*45. all*
6. to	*16. that*	*26. but*	*36. been*	*46. as*
7. is	*17. by*	*27. had*	*37. has*	*47. if*
8. was	*18. at*	*28. which*	*38. have*	*48. who*
9. it	*19. you*	*29. she*	*39. will*	*49. what*
10. for	*20. 's*	*30. they*	*40. would*	*50. said*

5.2.5 O trabalho de Stamatatos *et al* (2000b) para grego

Para seus experimentos com classificação de texto em grego segundo gêneros, Stamatatos *et al* (2000 b) criaram um *corpus* com textos da Web, sendo 25 textos de cada um dos seguintes 10 gêneros: editoriais, reportagens, textos científicos,

[41] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43

documentos oficiais, literatura, receitas, *curriculum vitae*, entrevistas, discursos e notícias pelo rádio (*broadcast news*).

Em seus experimentos são utilizadas 22 *features* (mostradas no Quadro 10), regressão múltipla (Aiken & West, 1996) e análise discriminante e 10 textos para treinamento e 10 para teste.

Quadro 10 – 22 features utilizadas nos experimentos para classificação de textos em gêneros de Stamatatos *et al* (2000 b)

1. detected sentences / words
2. punctuation marks / words
3. detected sentences / potential sentence boundaries
4. detected NPs / total detected chunks
5. detected VPs / total detected chunks
6. detected APs / total detected chunks
7. detected PPs / total detected chunks
8. detected CONs / total detected chunks
9. words included in NPs / detected NPs
10.words included in VPs / detected VPs
11.words included in APs / detected APs
12.words included in PPs / detected PPs
13.words included in CONs / detected CONs
14.detected keywords / words
15.special words / words
16.assigned morphological descriptions / words
17.chunks' morphological descriptions / total detected chunks
18.words remaining unanalyzed after pass 1 / words
19.words remaining unanalyzed after pass 2 / words
20.words remaining unanalyzed after pass 3 / words
21.words remaining unanalyzed after pass 4 / words
22.words remaining unanalyzed after pass 5 / words

Os agrupamentos detectados (*chunks*) podem ser: sintagmas nominais (NPs), sintagmas preposicionais (PPs), sintagmas verbais (VPs), e sintagmas adverbiais (APs). Dois agrupamentos são em geral conectados por uma seqüência de conjunções (CONs).

Os autores concluíram que as *features* 2, 12 e 15 são os marcadores de estilo mais importantes para a detecção de gênero para o problema tratado por eles. A taxa de acerto foi em média 82%, sendo que para alguns gêneros, como editoriais e *curriculum vitae*, foi em torno de 60%.

5.2.6 O trabalho de Dewdney *et al* (2001)

Dewdney *et al* (2001) utilizaram um *corpus* com 9705 textos com os gêneros: publicidade (1091), *bulletin board* (998), FAQs (1062), *message board* (1106),

notícias de rádio (2000), textos da agência de notícias Reuters (*Reuters newswire*) (2000) e notícias de TV (1448).

Sua lista de *features* contém 323 relacionadas à freqüência de palavras específicas e 89 outras *features* (nesse caso, marcadores de estilo) como número de adjetivos, proporção de mudanças no tempo verbal de um verbo para o próximo, freqüência de dias da semana e contagem de pontuação.

Em seus experimentos com os algoritmos *Naïve Bayes* (Mitchell, 2005), C4.5, e *Support Vector Machine* (SVM) (Cristianini & Shawe-Taylor, 2000), os autores constataram que, para os dois últimos algoritmos, o segundo conjunto de *features* tinha resultados melhores do que o primeiro que conta, apenas, com a freqüência de palavras. Concluíram que marcadores de estilo eram suficientes para a classificação de gêneros; que não é necessário o uso de *features* relacionadas a palavras específicas, quando se utiliza algoritmos de classificação apropriados. Segundo eles, os resultados ruins com marcadores de estilo e o algoritmo *Naive Bayes* estavam relacionados, por exemplo, ao fato desse algoritmo assumir independência entre as *features*, o que não é verdade, por exemplo, para a freqüência de etiquetas morfossintáticas. A melhor precisão é de 94,9% (+- 0,7%) com o algoritmo SVM e os dois conjuntos de *features* combinados.

5.2.7 O trabalho de Finn *et al* (2002)

Finn *et al* (2002) investigaram dois problemas de classificação em gêneros: (i) distinguir entre documentos objetivos e subjetivos; (ii) distinguir críticas positivas de críticas negativas. No primeiro caso, trabalharam com notícias que relatavam fatos/eventos objetivamente versus notícias que contêm a opinião do autor.

Para examinar o primeiro caso, construíram um *corpus* com 796 notícias extraídas da Web de três domínios: futebol, política, e finanças. Dessas, 422 relatavam fatos/eventos objetivamente e 374 continham a opinião do autor sobre o fato/evento. Para examinar o segundo caso, construíram um *corpus* com 1354 críticas de filmes e restaurantes extraídas automática e respectivamente dos sites Movie Review Query Engine[42] e Zagat[43]. Dessas, 686 eram positivas e 668 negativas.

[42] http://www.mrqe.com/lookup

Em ambos os casos são utilizados o algoritmo C4.5 e três conjuntos de *features*. Para compor o primeiro conjunto de *features* utilizam a estratégia *bag of words*, abordagem padrão na classificação de textos, em que cada documento é codificado como um vetor de *features*, sendo que cada elemento no vetor indica a presença ou ausência de uma palavra no documento. Nesse caso, utilizaram de *stemming* e remoção de *stopwords* para diminuir o tamanho do vetor. O segundo conjunto de *features* é composto por 36 etiquetas morfossintáticas (utilizam o etiquetador apresentado em Brill, 1994), representadas como uma porcentagem do número total de palavras de um texto. O segundo conjunto é composto por 152 *features*, sendo elas a freqüência de ocorrência de várias palavras funcionais e de símbolos de pontuação e estatísticas, como tamanho médio das frases, distribuição de palavras longas e tamanho médio das palavras. A taxa de acerto (*accuracy*) medida utilizando validação cruzada estratificada com 10 partes considerando-se apenas um dos domínios foi em média 87,2% utilizando *bag of words* para o primeiro problema e 82,7% para o segundo. Quando os testes são feitos com diferentes domínios a taxa de acerto média é de 78,5% com o uso de etiquetas morfossintáticas para o primeiro problema e 47,1% para o segundo problema. Na Tabela 5 mostramos as taxas de acerto médias.

Tabela 5 – Taxa de acerto para os dois problemas tratados por Finn *et al* (2002)

Bag of words	**Etiquetas morfossintáticas**	**Estatísticas**
Notícias que relatam fatos/eventos versus notícias que emitem opinião ***Treinamento e teste com mesmo domínio***		
87,2%	84,7 %	83,2%
Notícias que relatam fatos/eventos versus notícias que emitem opinião ***Treinamento e teste com domínios diferentes***		
82,7 %	61,3%	76,6%
Críticas positivas versus críticas negativas ***Treinamento e teste com mesmo domínio***		
67,3%	78,5%	67,8%
Críticas positivas versus críticas negativas ***Treinamento e teste com domínios diferentes***		
47,8%	47,1%	47,35%

Como podemos ver na Tabela 5, o uso do conjunto de *features* formado através da estratégia de *bag of words* é o melhor quando estamos restritos a um único domínio. Porém, quando os domínios de treinamento e teste são diferentes, o mesmo não acontece. Nesse caso, o melhor conjunto é o formado por etiquetas

[43] http://www.zagat.com/

morfossintáticas. Isto acontece porque, no primeiro caso, palavras que são referentes a um domínio específico são utilizadas como *features*. No caso do problema para diferenciar críticas positivas de negativas, por exemplo, a palavra "romântico", no caso de um restaurante, pode indicar uma crítica positiva, já, para filmes, a mesma palavra pode aparecer mais em críticas negativas do que positivas.

5.3 Considerações sobre a classificação em gêneros na busca diária de informação

Gêneros podem ser descritos como grupos de documentos que a) são estilisticamente consistentes e b) intuitivos para leitores experientes do canal de comunicação em questão (Kalrgren, 2004). Por exemplo, esse documento é um exemplo de monografia, tem uma forma que é familiar para pesquisadores (título, autor, resumo, introdução, seções de revisão da literatura, metodologia, resultados, conclusão) e deveria ter características estilisticamente consistentes com seu grupo – o de teses e dissertações (utilizar termos técnicos, ser formal, etc.).

Alguns gêneros são definidos primariamente em termos de propósito ou função, tais como proposta e requisição; outros, quanto à forma física, como brochura; outros, em termos da forma do documento, por exemplo, listas. Entretanto, a maioria dos gêneros implica uma combinação de propósito e forma, por exemplo, boletins informativos (*newsletter*), que é um conjunto de múltiplas notícias curtas distribuídas periodicamente para assinantes ou membros de uma organização. Dado que o gênero implica tanto forma como propósito, reconhecer o gênero de um documento a partir de sua forma pode prover informações sobre o propósito do documento.

Classificadores de textos quanto a gêneros, como os mostrados nos exemplos da Seção anterior, seriam intuitivamente muito úteis para a busca de informações na Web, para definirmos se um documento trata do assunto que estamos interessados com o enfoque desejado. Entretanto, há quatro problemas relacionados à aplicação da maioria desses estudos em máquinas de busca na Web, todos relacionados ao *corpus* utilizado para treinamento dos classificadores: (i) *corpus* como o Brown não foram feitos com o propósito de serem utilizados para detecção de gêneros, por isso nem sempre são estilisticamente homogêneos (Kessler *et al*, 1997); (ii) alguns dos *corpora*

são muito homogêneos quanto a suas fontes, como no caso dos formados apenas por notícias ou textos técnicos; (iii) qual a utilidade prática de categorias como "curiosidades/fábulas/folclore/cultura popular" (*Popular Lore*) (Kessler *et al*, 1997); (iv) quais os gêneros presentes na Web, quais gêneros novos foram criados e são importantes, os gêneros estudados anteriormente em outras mídias têm as mesmas características na Web? Gêneros são dependentes de contexto, e o leque de escolhas de um jornal, por exemplo, é diferente do leque do clube do livro.

De acordo com as respostas dadas para o questionário mostrado no Apêndice B e que são analisadas no Capítulo 7, acreditamos que o reconhecimento automático de gêneros, considerando-se o conceito de gêneros de Karlgren mencionado acima, pode facilitar muitas das buscas na Web, em sistemas nos quais os usuários poderiam especificar os gêneros de interesse para eles ou as tarefas que estão tentando realizar. É importante para isso que se escolha bem os gêneros que serão tratados direta ou indiretamente, já que tratar especificamente todos os gêneros que possam existir na Web é uma tarefa impossível, pois os gêneros variam dentre outras coisas de acordo com a cultura e o tempo. Também, que utilizemos em nossos treinamentos textos da Web, dos mais diversos domínios, fontes, autores, etc.

No próximo capítulo apresentamos: (i) os *corpora* utilizados nesse trabalho de doutorado, discutindo os gêneros e tarefas/necessidades de busca tratados; (ii) os conjuntos de marcadores de estilo; (iii) o protótipo de ferramenta de busca desenvolvido para que pudéssemos testar a nossa hipótese de que a classificação de necessidades poupa esforços na identificação de documentos relevantes; e (iv) os três algoritmos utilizados na maioria dos experimentos: J48, SMO e LMT.

6. Classificação automática de resultados segundo a intenção de busca

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 form[44].

Neste capítulo, apresentamos os *corpora*, os conjunto de marcadores e os algoritmos utilizados nos experimentos para criação de classificadores de textos segundo gêneros, tipos textuais, necessidades gerais (sete necessidades) e necessidades personalizadas (relatados nos Capítulos 8 e 9). Também apresentamos o protótipo de um meta-buscador criado para que a classificação em necessidades gerais e personalizadas pudessem ser avaliadas com usuários (Capítulos 9 e 10).

6.1 Modos de classificação explorados neste trabalho

O objetivo desta pesquisa foi encontrar formas de apresentar resultados de sistemas de RI da Web que considerassem, além do tópico do documento, o enfoque que o usuário espera que o documento aborde sobre um assunto. Nossa hipótese inicial foi que marcadores de estilo, como os utilizados na classificação de textos em gêneros (apresentados no Capítulo 5), poderiam ser utilizados para classificar textos segundo necessidades gerais de busca (sete necessidades mostradas na Seção 6.1.3), uma vez que tais necessidades são, em geral, melhor atendidas por tipos de texto específicos. Porém, a idéia de Biber (1995) de que marcadores semelhantes aos utilizados por ele poderiam ser utilizados também em outras línguas para a classificação de textos, não havia sido testada para português. Como mostrado no Capítulo 5, em nossa revisão bibliográfica sobre classificação automática segundo gêneros, utilizando marcadores de estilo, encontramos apenas trabalhos para inglês, grego e alemão.

Nossos primeiros experimentos (Aires *et al,* 2004a, 2004b) foram encorajadores (os resultados são mostrados no Capítulo 8), mas não foram bons para o esquema de classificação em sete necessidades como um todo. Uma outra investigação que se fez necessária foi verificar se os marcadores de estilo por nós

[44] *http://www.dcs.shef.ac.uk/research/groups/nlp/extraction*

adaptados para o português do trabalho de Biber (1988) e de Karlgren (2000), realmente funcionavam para a classificação de textos em gêneros e tipos textuais em português, o que foi verificado em nossos experimentos para a tipologia textual do *corpus* Lácio-Ref do Projeto Lácio-Web[45], descrita nas Seções 6.1.1 e 6.1.2.

Ao longo dos experimentos que são apresentados no Capítulo 8, nos quais verificamos nossa hipótese inicial citada acima, pensou-se em classificar os textos segundo eixos mais flexíveis do que as sete necessidades, nos quais um texto poderia ser: (i) formal ou informal; (ii) contextualizado ou não; (iii) apenas descritivo ou emitir opinião. Porém, durante a criação de um *corpus* de treinamento com textos classificados segundo esses critérios, verificamos que os critérios, aparentemente tão simples, podiam ser interpretados de formas diferentes pelos usuários (mais detalhes são dados na Seção 6.1.3). Como alternativa para esse esquema de três eixos, pensamos em nossa segunda hipótese, a de que os marcadores de estilo poderiam ser também utilizados para criação de esquemas de classificação binários de necessidades personalizadas (descritos na Seção 6.1.4).

Em suma, neste trabalho, o enfoque desejado pode ser selecionado de uma taxonomia de gêneros (Seção 6.1.1), de tipos textuais (Seção 6.1.2), de necessidades de busca (Seção 6.1.3) ou de taxonomias binárias de necessidades personalizadas (Seção 6.1.4). Nas próximas subseções, mostramos os *corpora* utilizados na criação de classificadores para cada um desses quatro tipos de enfoque, mostrados em ordem do enfoque mais genérico ao mais específico.

6.1.1 Gêneros

Para a classificação em gêneros, utilizamos os gêneros do *corpus* Lácio-Ref (Aluísio *et al*, 2003). O Lácio-Ref é um *corpus* aberto e de referência do português contemporâneo do Projeto Lácio-Web, composto de textos em português brasileiro, tendo como característica serem escritos respeitando a norma culta. A taxonomia de gêneros utilizada para classificar os textos do *corpus* pode ser vista no Quadro 11.

[45] http://www.nilc.icmc.usp.br/lacioweb/

Quadro 11 - Taxonomia de gêneros do Lácio-Ref

Gêneros	Exemplos de textos
Científico	Artigo, tese, projeto...
De referência	Enciclopédia, dicionário, glossário,...
Informativo	Reportagem, notícia,...
Jurídico	Lei, sentença, medida provisória,...
Prosa	Biografia, conto, novela, romance e outros.
Poesia	
Drama	
Instrucional	Livro-texto, receita culinária, apostila...
Técnico-Administrativo	Carta, memorando, manual...

Entretanto, em sua versão atual, o *corpus* não contém textos do gênero de referência ou do gênero técnico-administrativo. Em nossos experimentos com classificação em gêneros, utilizamos os textos dos gêneros disponíveis e reunimos os gêneros poesia, prosa e drama em um único supergênero Literário. A Tabela 6 mostra o número de textos de cada gênero.

Tabela 6 - Número de textos por gênero do Lácio-Ref

Gêneros	Número de textos
Científico	202
Informativo	3792
Jurídico	49
Literário	232
Instrucional	3

Como o gênero Instrucional continha apenas 3 textos, selecionamos mais 280 textos da Web dentre receitas, apostilas e manuais.

Como o gênero Informativo é representado no *corpus* por textos jornalísticos e esses têm estilo diferente na Web, substituímos os 3792 textos jornalísticos originais do *corpus* por 150 textos extraídos da Web.

6.1.2 Tipos Textuais

Para a classificação em tipos textuais também utilizamos o *corpus* Lácio-Ref (Aluísio *et al*, 2003). Os tipos de texto considerados no Lácio-Ref podem ser vistos no Quadro 12. É importante dizer que apenas os tipos que aparecem negritados no Quadro 12 (28 tipos) têm textos na versão atual do Lácio-Ref.

Apostila	**Declaração** 2 108	Manual	**Petição** 0 199	**Reportagem** 2165 2171
Artigo 363 363	**Decreto** 3 12	**Medida Provisória** 178 178	**Poema** 21 146	**Resenha** 56 160
Ata	**Edital** 4 227	Memorando	**Portaria** 3 415	**Resolução** 3 243
Boletim	**Editorial** 44 119	**Monografia** 46 46	Projeto	**Resumo** 30 107
Carta 13 13	Ensaio	Notas Didáticas	**Provimento** 3 3	**Sentença** 6 6
Circular 3 58	**Entrevista** 108 108	**Notícia** 1096 1096	**Receita** 1 189	**Súmula** 5 180
Contrato	**Lei** 7 179	**Ofício** 3 44	**Regimento** 3 3	Testamento
Crônica 62 228	**Livro-Texto** 186 186	**Parecer** 3 186	**Relatório** 28 28	Verbete

Quadro 12 – Tipos textuais do Lácio-Ref

Como alguns tipos textuais continham um número muito pequeno de textos, incluímos mais alguns textos retirados da Web. O número original de textos por tipo é o primeiro no Quadro 12 e o segundo número é o número de textos na versão do *corpus* com textos da Web. O acréscimo de textos foi realizado somente para algumas entradas de forma não sistematizada, sem o intuíto de balancear o corpus. A versão original do *corpus* possui 28 tipos de texto, num total de 4445 textos. A versão aumentada possui 29 tipos, num total de 7001 textos.

6.1.3 Necessidades de busca

O esquema de classificação em necessidades de busca é resultado de uma análise qualitativa, realizada pela autora desta tese, dos *logs* de novembro de 1999 e de julho de 2002 da máquina de busca TodoBr[46]. A interpretação de quais necessidades estão por trás de uma dada consulta não é uma tarefa exata, mesmo quando estão disponíveis dados como que *links* o usuário visitou e quanto tempo demorou para

46 Principal máquina de busca do domínio .br que foi recentemente (em julho de 2005) incorporada ao Google, como Google Brasil: http://www.todobr.com.br/

retornar a lista de resultados apresentada pela máquina de busca, pois o usuário pode ter verificado após o clique que o resultado não era relevante para sua consulta. Duas estratégias alternativas para identificar quais necessidades de busca tratar seriam: (i) perguntar a pessoas em geral por *e-mail* quais necessidades deveríamos tratar; (ii) tendo-se uma máquina de busca disponível, mostrar aos usuários, quando fizessem uma consulta, uma tela de *feedback* na qual descreveriam o objetivo por trás de uma dada consulta. No nosso caso, nos restaria a primeira opção acima, porém, sabe-se que o número de respostas é pequeno quando não se tem uma recompensa imediata a oferecer. Veja, por exemplo, o caso de Karlgren (2000) (67 pessoas de 648 questionadas responderam) e Aires & Aluísio (2003) em que 16 pessoas responderam de 440 questionadas. Por isso, por podermos cobrir um número maior de interesses, a análise de *logs*, apesar de não ser totalmente precisa, era a melhor alternativa.

Os sete tipos de necessidades com os quais trabalhamos são encontrar:

1 – Páginas que definam alguma coisa ou ensinem como e/ou porque algo acontece. Por exemplo: o que é a aurora boreal. Para esta necessidade, os melhores resultados seriam dicionários e enciclopédias, livros didáticos, artigos técnicos e relatórios e textos do gênero informativo.

2 – Páginas que ensinem como fazer algo ou como algo é feito. Por exemplo: instruções de como instalar Linux em seu computador, receita de um bolo. Resultados típicos seriam textos do gênero instrucional, tais como manuais, livros didáticos, receitas e também alguns artigos técnicos e relatórios.

3 – Páginas que forneçam uma apresentação (ou apanhado ou panorama) sobre um determinado assunto. Por exemplo, um panorama sobre a literatura americana no século XX. Nesse caso, os melhores textos seriam dos gêneros instrucionais, informativo e científico, por exemplo, reportagens.

4 – Páginas com notícias. Por exemplo: uma notícia sobre um atentado. As melhores respostas seriam textos do gênero informativo, como, por exemplo, notícias em jornais e revistas.

5 – Páginas que forneçam informações sobre uma pessoa, ou empresa, ou instituição, ou organização. Por exemplo: páginas pessoais, páginas com informações

para contato (com currículo, telefone, endereço). Respostas típicas seriam páginas pessoais e institucionais.

6 – Uma página específica que o usuário quer visitar, mas não se lembra da URL. Nesse caso, os resultados poderiam ser de qualquer tipo textual ou gênero.

7 – Páginas que forneçam algum serviço online. Por exemplo: lojas virtuais, serviço dos correios para acompanhamento de envio de encomendas. As melhores respostas, nesse caso, seriam textos comerciais (empresas ou indivíduos oferecendo produtos e serviços).

Sabemos que usuários podem fazer todo tipo de consultas e que por isso as sete necessidades acima provavelmente não atendem a todo usuário em todas suas buscas. Entretanto, é importante ressaltar que em nossa taxonomia cobrimos as necessidades informacional, navegacional e transacional, que são citadas por diversos autores como os três tipos básicos de consultas (Broder, 2002; Kang & Kim, 2003; Rose & Levinson, 2004).

Para criar o *corpus* de textos classificados segundo as sete necessidades acima, poderíamos ter seguido duas estratégias: (i) analisar as páginas de coleções de páginas da Web, como a WBR-99 (Calado, 1999); ou (ii) selecionar páginas de locais da Web que tipicamente contêm páginas que atendem aos tipos de necessidades descritos acima. Para poupar tempo, utilizamos a segunda estratégia. A primeira versão do *corpus* foi criada por cinco pessoas, cada uma responsável por coletar páginas de um dos cinco tipos de necessidades, seguindo os seguintes critérios: (i) as páginas deveriam ser escritas em português do Brasil, para que variações lexicais, morfológicas e sintáticas entre as diversas variantes não interferissem no treinamento dos classificadores e (ii) as páginas selecionadas deveriam ser de diversas fontes e assuntos, já que, como visto no Capítulo 5, textos de uma mesma fonte ou área podem ter estilo próprio (por exemplo, textos da Folha de São Paulo e textos médicos), sendo que o que pretendíamos investigar eram os marcadores de estilo relacionados ao propósito do texto. Após a seleção, todos os textos foram revistos pela autora desta tese para verificar se os critérios definidos haviam sido atendidos.

O *corpus* foi criado inicialmente com 511 textos, 73 de cada tipo de necessidade (com exceção do tipo 6) e 73 adicionais que não atendem a nenhum dos

tipos de necessidades (que chamamos de "outros"). Como pôde ser visto no Capítulo 5, os trabalhos sobre classificação automática de gêneros não seguem padrões; são estudos empíricos em que o número de *features* e tamanho do *corpus* varia muito. Para decidir qual deveria ser o tamanho do *corpus*, seguimos a recomendação de Gorsuch (1983:332, *apud* Biber 1988:65) de que os dados para a análise fatorial (*factor analysis*)[47] devem incluir cinco vezes mais textos do que o número de características lingüísticas a serem analisadas. Apesar de estarmos fazendo outro tipo de análise, seguimos essa recomendação (nosso conjunto original de *features* era composto por 46 marcadores de estilo, como será explicado na Seção 6.3).

Os 511 textos foram convertidos para o formato texto, mantendo-se todo o texto de páginas, inclusive o contido em *frames* (perdemos apenas textos que faziam parte de figuras), mas *links* não foram seguidos. O *corpus* resultante contém 640.630 palavras, o número de palavras por tipo de necessidade pode ser visto na Tabela 7.

Tabela 7 – Número de palavras por necessidade da primeira versão do *corpus* de necessidades

Tipo 1	Tipo 2	Tipo 3	Tipo 4	Tipo 5	Tipo 7	Outros
76.841	51.959	196.450	39.533	67.601	39.951	168.295

Nessa primeira versão do *corpus*, não foi considerado o fato de que um mesmo texto pode atender a mais de uma necessidade. Por isso, todos os textos foram revistos para que pudéssemos reclassificá-los considerando esse fato. Nessa revisão encontramos 22 combinações dos tipos de necessidades tratados: 1, 2, 3, 4, 5, 7, Outros, 1-2, 1-3, 1-4, 2-3, 2-4, 2-7, 3-4, 3-5, 3-7, 4-5, 1-2-3, 1-3-4, 1-3-5, 1-3-7 e 1-2-3-7. Nosso próximo passo foi aumentar o número de textos para cada um dos tipos encontrados. O número de textos e palavras para cada tipo é mostrado na Tabela 8[48].

As recomendações gerais para a reclassificação e aumento do *corpus* foram:

- escolha apenas páginas em português do Brasil;

julgue todo o conteúdo da página, não apenas títulos ou o que apareça em destaque;

- não considere o gênero do texto, o que importa é que satisfaça a necessidade de busca;

[47] http://www2.chass.ncsu.edu/garson/pa765/factor.htm
[48] Essa versão do *corpus* se encontra disponível em www.linguateca.pt/Repositorio/yesuser.html

- não se preocupe com a qualidade ou quantidade de informação do texto: se uma página tem pouca informação para satisfazer ao Tipo A e muita informação para satisfazer o tipo B, deve ser classificada como AB;
- classifique apenas a página, não siga *links* para outras páginas. Por exemplo, se uma página tem um link que diz "clique aqui para se inscrever", essa página não fornece um serviço de inscrição. Mas se a página tem um link indicando "Download" para que se faça download de algo, essa página oferece um serviço. Além disso, foram também dadas instruções sobre cada tipo de necessidade, com exemplos e contra-exemplos de textos (Aires *et al*, 2005b).

Tabela 8 – Número de textos e palavras na versão final do *corpus* de necessidades

Necessidades	Textos	Palavras
T1	78	61.068
T12	77	102.008
T123	77	122.003
T1237	80	69.999
T13	72	134.954
T134	80	89.999
T135	79	64.656
T137	79	51.925
T14	80	56.270
T2	77	72.303
T23	76	88.724
T24	79	72.791
T27	79	44.497
T3	77	149.387
T34	75	81.785
T35	79	77.767
T37	80	52.943
T4	75	87.030
T45	79	64.799
T5	69	65.852
T7	77	53.639
Outros	76	137.563
Total	1.703	1.801.962

6.1.4 Necessidades de busca personalizadas

Como dito anteriormente, os sete tipos de necessidades descritos na Seção 6.1.3 não cobrem todos os tipos de busca que um usuário possa vir a fazer. Por isso, a princípio pensamos também em oferecer uma forma de classificação mais flexível, baseada no

fato de o usuário querer um texto formal ou informal, contextualizado ou não e que emitisse opinião ou fosse apenas descritivo. Porém, durante a criação de um *corpus* de treinamento com textos classificados segundo esses três critérios, verificamos que os critérios, aparentemente tão simples, podiam ser interpretados de formas diferentes pelos usuários.

Por exemplo, o que seria um texto formal? Um texto formal é um texto que segue a norma culta da língua? Um texto formal deveria seguir a norma culta da língua, mas não podemos garantir que todos sigam. O grau de formalidade (superformal, formal, semiformal ou informal) depende na verdade da circunstância em que o texto será empregado. No caso da nossa classificação de textos da Internet, alguns exemplos seriam: do superformal – pareceres judiciais; do formal – um verbete; do semiformal – uma notícia em jornal; do informal – uma opinião em um blog.

A subjetividade por trás desses três critérios foi confirmada durante a reclassificação do *corpus* de necessidades segundo esses três eixos. Os textos deveriam ser classificados por duas pessoas. Ambas receberam as definições de cada um dos critérios e treinamento através do acompanhamento de seus trabalhos por uma semana. Após o treinamento, os 22 grupos de textos foram divididos entre ambas; uma das pessoas foi responsável por classificar textos de 11 grupos e outra por classificar textos de 12 grupos (para que um dos grupos fosse feito por ambas). Mesmo com treinamento e com acompanhamento constante durante todo o processo, a classificação dos textos do grupo em comum teve 30% de divergência. Todos os textos foram revistos pela autora desta tese, que identificou ainda inconsistências na classificação dos textos de uma mesma pessoa, que, para textos diferentes, pareceu tomar diferentes estratégias. Em entrevista posterior, quando questionadas sobre as divergências, ambas definiram os critérios da mesma forma que inicialmente. Suas justificativas, para escolhas diferentes para o que pareciam textos de um mesmo tipo, foram relacionadas a características particulares dos textos, que não poderiam ser utilizadas facilmente para tecer recomendações gerais. Por exemplo, justificar que um texto era formal devido a certas escolhas lexicais do autor, que, na verdade, estavam relacionadas ao assunto do texto. Ambas as pessoas responsáveis pela classificação

têm formação em Letras, enquanto a pessoa responsável pelo treinamento e acompanhamento tem formação em Computação, subárea PLN.

Dada a subjetividade dos critérios acima mencionados, optamos por fornecer ao usuário uma possibilidade de classificação ainda mais personalizada, permitindo que o usuário crie seu próprio esquema de categorização. Nessa opção, o usuário fornece exemplos de textos de um problema com o qual lida freqüentemente em suas buscas na Internet, e o sistema, através de marcadores estilísticos, gera um esquema de classificação novo para aquele usuário. Os exemplos devem ser de problemas binários (de duas classes), que estejam relacionados a tipos de texto, assim como as sete necessidades citadas anteriormente. Por exemplo, no caso de um advogado, distinguir entre textos técnicos sobre direito e textos voltados para o público comum, como a sentença dada para uma determinada ação e uma página informal sobre os direitos do consumidor, respectivamente. Essa abordagem não serve para problemas de classificação relacionados ao assunto, como, por exemplo, distinguir entre textos científicos que falam sobre problemas do coração da área de cardiologia e textos de outras áreas médicas que também falem sobre problemas do coração.

Para testar a opção personalizada utilizamos inicialmente dois *corpora*, um criado em um mestrado do ICMC (Martins & Moreira, 2004), e outro criado por nós, para testes. O propósito do primeiro *corpus* é distinguir se uma página contém descrições de produtos à venda ou não e é composto por 1252 páginas (723 exemplos positivos e 529 negativos). O segundo *corpus* é composto por 200 páginas relacionadas ao domínio de direito; tem o propósito de distinguir entre páginas de direito para pessoas da área (advogados, juízes, etc.) e textos para pessoas em geral, sendo formado por 100 exemplos positivos e 100 negativos.

Essa opção de classificação foi testada também com sete *corpora* criados por usuários para seus problemas específicos. Esses *corpora* são descritos no Capítulo 9.

6.2 Algoritmos

Em nossos primeiros experimentos (Aires *et al*, 2004 b), utilizamos o algoritmo de classificação J48, disponibilizado na coleção de algoritmos de aprendizado de máquina Weka (Witten & Frank, 2000). J48 é a versão do Weka do algoritmo C4.5. C4.5 foi escolhido por ser um algoritmo bem conhecido e bastante discutido, por ter

sido utilizado em estudos similares (Karlgren, 2000), e por gerar regras de fácil compreensão.

Entretanto, não existem comparações amplas de algoritmos de classificação para uso com marcadores estilísticos, como existem para outros domínios. Por isso, experimentamos todos os algoritmos disponíveis no Weka que pudessem lidar com *features* numéricas, com classes não numéricas e com o número de classes que precisávamos. Utilizamos um total de 44 algoritmos (Aires *et al*, 2004a): *Naive Bayes, Naive Bayes Multinomial, Naive Bayes Updateable, Multilayer Perceptron, SMO, Simple Logistic, IB1, IBK, KStar, LWL , AdaBoostM1, Attributive Selected Classifier, Bagging, Classification via regression, CV parameter selection, Decorate, Filtered classifier, Logit Boost, Multiclass classifier, Multi Scheme, Ordinal class classifier, Raced incremental logit boost, Random committee, Stacking, Stacking C, Vote, FLR, HyperPipes, VFI, Decision Stump, J48, LMT, Random Forest, Random Tree, REP Tree, User classifier, ZeroR, Conjunctive Rule, OneR, Decision Table, Part, NNGe, Ridor,* e *JRIP*. Desses, selecionamos três para utilizar nos experimentos relatados nos próximos Capítulos: J48, Sequential Minimal Optimization (SMO) e Logistic Model Tree (LMT).

O método C4.5 (Quinlan, 1993) ordena os fatores (features) durante o treinamento utilizando o algoritmo de ganho de informação (*informaton gain*): constrói uma árvore cujos nós representam decisões que dividem os dados em dois grupos, usando, de todos os fatores ainda não considerados, o que leva a um ganho maior. As folhas da árvore representam pontos em que uma classificação é atribuída. A árvore é então podada substituindo subárvores por folhas, se essa substituição reduzir o erro esperado, o que minimiza a adaptação aos erros (*error fitting*) e reduz a complexidade da árvore. A nova árvore é então o classificador, pronto para ser utilizado sobre novos dados. A classificação procede então do seguinte modo: em cada nó aplica-se a regra, calculando o respectivo fator do documento a classificar, o que leva a uma nova regra. O documento fica classificado quando o algoritmo chega a uma folha.

SMO (*sequential minimal optimisation*) é a implementação do Weka para algoritmo de otimização do treino de classificadores do tipo "*support vector machine*" (SVM) usando kernels polinomiais ou RBF de Platt (1998). Esse algoritmo

transforma a saída do classificador SVM em probabilidades através da aplicação de uma função sigmóide padrão que não é adaptada (*fitted*) aos dados. Essa implementação não é rápida com um espaço de fatores linear, nem com dados esparsos. Substitui todos os valores que faltarem, transforma atributos nominais em binários e normaliza todos os valores numéricos.

O LMT (Landwehr *et al*, 2003) é um algoritmo para classificação que constrói *logistic model trees*, que são árvores de classificação com funções de regressão logística em suas folhas.

6.3 Marcadores estilísticos

Como mostrado no Capítulo 5, a seleção inicial de marcadores estilísticos é uma tarefa empírica. Por isso, nessa tese, experimentamos diferentes conjuntos de marcadores estilísticos[49]. O primeiro, que é utilizado na maioria dos experimentos, foi criado com base nos trabalhos de Biber (1988) e Karlgren (2000). Para sua criação levamos em consideração nossas intuições lingüísticas e demos preferência a marcadores que pudessem ser calculados sem a ajuda de qualquer tipo de analisador, como etiquetadores morfossintáticos e sintáticos. Esse conjunto contém 46 marcadores e pode ser visto no Quadro 13. Referir-nos-emos a eles como marcadores baseados em estatísticas.

Quadro 13 – 46 marcadores estilísticos utilizados como *features* nos primeiros experimentos de classificação (Aires *et al* 2004a, 2004b)

Estatísticas baseadas em palavras
Estimativa de itens lexicais diferentes, dado pelo número de itens lexicais diferentes dividido pelo número de itens (*type/token ratio*)
Estimativa de itens lexicais diferentes, porém considerando-se apenas os itens iniciados por letra maiúscula (*capital type token ratio*)
Número de dígitos
Tamanho médio das palavras em caracteres
Número de palavras longas (com mais de 6 caracteres)
Estatísticas baseadas no texto como um todo
Número de caracteres
Tamanho médio das frases em caracteres
Número de frases
Tamanho médio das frases em palavras
Tamanho do texto em palavras
Outras estatísticas.
Número de ocorrências das expressões “acho”, “acredito que”, “parece que” e “tenho impressão

[49] Todos os marcadores estilísticos utilizados neste projeto, que sejam freqüências simples, são normalizados para verificarmos suas freqüências a cada 100 palavras; isto é feito devido ao fato de que os tamanhos de texto variam muito e grande parte dos textos menores tem cerca de 100 palavras.

que”
As palavras “é” e “são”
A palavra “que”
A palavra “se”
Os marcadores discursivos “agora”, “da mesma forma”, “de qualquer forma”, “de qualquer maneira” e “desse modo”
As palavras “aonde”, “como”, “onde”, “por que”, “qual”, “quando”, “que” e “quem” no início de perguntas
“E”, “ou” e “mas” no início de frases
Amplificadores (*amplifiers*) (Quirk *et al*, 1992). Alguns exemplos são: “absolutamente”, “extremamente” e “completamente”.
Conjuncts. (Quirk *et al*, 1992). Alguns exemplos são: “além disso”, “consequentemente”, “assim” e “entretanto”.
Downtoners (Quirk *et al*, 1992). Alguns exemplos são: “com exceção”, “levemente”, “parcialmente” e “praticamente”.
Enfáticos (*emphasizers*) (Quirk *et al*, 1992). Alguns exemplos são: “definitivamente”, “é óbvio que”, “francamente” e “literalmente”.
Verbos suasivos (Quirk *et al*, 1992) como aderir, crer e dar
Verbos privados (Quirk *et al*, 1992) como ter e guardar
Verbos públicos (Quirk *et al*, 1992) como abolir, promulgar e mencionar
Número de artigos definidos Número de artigos indefinidos
Pronomes na primeira pessoa Pronomes na segunda pessoa Pronomes na terceira pessoa
Número de pronomes demonstrativos
Pronomes indefinidos e expressões pronominais
Número de preposições
Advérbios de lugar Advérbios de tempo
Número de advérbios
Número de interjeições
Contrações
Conjunções causais Conjunções finais Conjunções proporcionais Conjunções temporais Conjunções concessivas Conjunções condicionais Conjunções conformavas Conjunções comparativas Conjunções consecutivas

O segundo conjunto de marcadores utilizado é composto por cinco funções para medir a riqueza de vocabulário (Stamatatos *et al*, 2000b):

$$K = \frac{10^4 (\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2} \qquad W = N^{V^{-\alpha}} \qquad R = \frac{(100 \log N)}{\left(1 - \left(\frac{V_1}{V}\right)\right)}$$

$$D = \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)} \qquad S = \frac{V_2}{V}$$

onde:

Vi é o número de palavras utilizadas i vezes

α é igual a 0.17

Referir-nos-emos a esse conjunto como marcadores de riqueza vocabular.

O terceiro conjunto de marcadores é composto pelas 62 palavras mais freqüentes do *corpus* de necessidades: eliminando-se as *stopwords*, verbos auxiliares, advérbios, palavras relacionadas a domínios e agrupando algumas das palavras mais freqüentes como um único marcador. Chamamos esse conjunto de características léxicas; os 62 marcadores são mostrados no Quadro 14. Esse conjunto é utilizado apenas nos experimentos com a classificação em sete necessidades de busca, pois possui marcadores dependentes dessa tarefa, como, por exemplo, número de ocorrência das palavras ”download” e “kb”.

Quadro 14 – 62 marcadores selecionados a partir da análise das palavras mais freqüentes do *corpus* de necessidades

1. área	17. empresa, grupo	33. notícias	49.reservados
2. arquivo, arquivos	18. estado	34. número	50.segue, seguem, siga, sigam
3. busca	19. exemplo	35. página, site	tem, têm
4. clique	20. fale	36. país	51.serviços
5. como	21. faz, faça, façam	37. partir	52.sistema
6. contato, mail	22. fazer	38. passos	53.tecnologia
7. copyright	23. forma	39. pesquisa	54.reservados
8. dados	24. geral	40. pessoa, pessoas	55.tempo
9. define, definem, definiu, definiram	25. home	41. pode, podem	56.texto
10. deve, devem	26. http, www	42. política	57.tipo
11. dia, dias	27. importante	43. porquê	58.uso
12. direitos	28. informações	44. processo	59.ver
13. diz, dizem, conta contam, relata, relatou, disse	29. internet, online	45. produção	60.versão
14. download, kb	30. mercado	46. produto, produtos	61.vez
15. foi, foram, será, serão	31. nacional	47. qualidade	62.vezes
16. edição	32. nome	48. referências	

Além dos três conjuntos de marcadores já mencionados, fizemos uso também de dois outros conjuntos, em experimentos na classificação segundo necessidades: marcadores sintáticos e marcadores de aparência gráfica (*layout*).

São 15 os marcadores sintáticos selecionados com base em nossa intuição lingüística, que foram calculados com ajuda do etiquetador sintático PALAVRAS

(Bick, 2000)[50]. De acordo com o autor e como relatado em Bick (2000), o PALAVRAS tem um desempenho geral de 99% para atribuição de etiquetas morfossintáticas e 95-96% para atribuição da função sintática. Porém não avaliamos a taxa de acerto para o *corpus* de necessidades. A lista de marcadores pode ser vista no Quadro 15.

Quadro 15 – 15 Marcadores estilísticos sintáticos

Tipos de sujeito
1. Porcentagem de sujeitos que são sintagmas nominais
2. Porcentagem de sujeitos pronominais
3. Porcentagem de sujeitos que são nomes próprios
Sobre os nomes próprios
4. Número de nomes próprios
5. Tamanho médio dos nomes próprios em palavras
Sobre os verbos
6. Número de verbos
7. Número de verbos na primeira pessoa
8. Número de verbos na segunda pessoa
9. Número de verbos na terceira pessoa
10. Porcentagem de verbos auxiliares
11. Porcentagem de verbos principais
12. Porcentagem de verbos no infinitivo
13. Porcentagem de verbos no gerúndio
14. Número de verbos/número de sentenças
Sobre as orações subordinativas
15. Número de orações subordinadas (finitas, infinitas e adverbiais)

O último conjunto de marcadores é composto por 27 marcadores de aparência gráfica. O que chamamos de marcadores de aparência gráfica são características de documentos html que indicam decisões como fonte utilizada, espaçamento e cor. Os 27 marcadores utilizados podem ser vistos no Quadro 16.

Quadro 16 – 27 marcadores estilísticos baseados em característica da aparência gráfica de documentos

Tags relacionadas à estrutura dos documentos
1. <Hn n vai de 1 a 6 Para especificar títulos em ordem de importância (de 1 a 6)

[50] A etiquetação do *corpus* de necessidades com o PALAVRAS levou um dia para que fosse concluída em um Pentium 4 - 2,4 Ghz, com 1 GB de RAM, HD de 28 Gb e SO = Red Hat Linux 9.

2. <ADDRESS Para fornecer informação de contato sobre o autor da página
3. <P> Parágrafo
4. <UL> e <OL> Informações apresentadas em listas (ordenadas ou não)
5. <DL> Lista de definições para termos
6. <DIV Para especificar divisão em blocos do documento
7. <BLOCKQUOTE> Citações
8. <FORM Formulários, tais como de compras e registro
9. <HR Linha para indicar mudança de tópicos, separar partes dos documentos
10. <TABLE informação em tabela
Tags relacionadas a cores
11. BGCOLOR= e TEXT= e LINK= e VLINK= e ALINK= Cor do fundo do documento, do texto e de *links*
Marcadores de frase e fonte
12. <EM>, <STRONG>, <DFN>, <CODE>, <SAMP>, <KBD>, <VAR>, <CITE>, TT, I, B, U, STRIKE, BIG, SMALL, SUB, SUP e <FONT e <BASEFONT Negrito, itálico, tamanho da letra etc.=
Campos de entrada
13. <INPUT TYPE=TEXT campo de texto
14. <INPUT TYPE=PASSWORD campo de senha
15. <INPUT TYPE=CHECKBOX campo de atributos booleanos
16. <INPUT TYPE=RADIO para atributos que podem ter apenas um valor, dado um conjunto de valores
Outras tags
17. <ISINDEX Entrada de palavras-chave para busca simples
18. <BASE Para definir url base para url relativas no documento
19. <IMG Para incluir figuras no documento
20. <META Para dar mais informações sobre os documentos, como descrição, palavras-chave, etc.
21. <LINK Para especificar relações com outros documentos
22. <SCRIPT

23. <STYLE Para especificar style sheets para utilizar com um documento
24. <A HREF= + <A NAME= hiperlinks e ancoras
25. <APPLET Para incluir java applets no documento
26. <MAP Image maps
27. <SELECT Para especificar um menu de alternativas a serem selecionadas

6.4 Leva-e-traz

Para que pudéssemos testar os esquemas de classificação criados com usuários, construímos o Leva-e-traz[51], um protótipo de meta buscador para português. Escolhemos desenvolver um meta-buscador para que não tivéssemos que implementar a indexação e a correspondência entre documentos e consultas.

O Leva-e-traz executa as consultas digitadas pelo usuário no Google e no AlltheWeb; pega os primeiros 10 resultados de cada uma das máquinas de busca e elimina os resultados repetidos. A decisão de retornarmos apenas 20 resultados foi tomada por questões de tempo. O download das páginas que precede o cálculo das *features* encarece o processo de classificação, fazendo com que o tempo de resposta possa chegar a pouco mais de três minutos. O Leva-e-traz classifica os resultados por gêneros, tipos textuais, necessidades de busca e necessidades personalizadas. Permite, também, a criação de necessidades personalizadas. Através de sua tela principal, é possível também obter resultados sem qualquer classificação. Isso foi feito para que pudéssemos comparar o esforço de busca dos usuários, dadas a apresentação dos resultados tradicional e a apresentação dos resultados classificados segundo as sete necessidades de busca; essa avaliação será descrita no Capítulo 10. Todas as telas do Leva-e-traz podem ser vistas no Apêndice A. As classificações em gêneros e em tipos textuais foram incluídas no Leva-e-traz por terem sido consideradas úteis por usuários na avaliação que será descrita no próximo Capítulo 7; porém, testes com usuários foram feitos apenas para a classificação em necessidades e a classificação em necessidades personalizadas (Capítulo 9).

A interface do Leva-e-traz foi construída usando a linguagem Flash. A busca é feita utilizando-se uma API[52] disponibilizada pelo Google. Posteriormente, faz-se download dos resultados que são, então, transformados em txt utilizando-se um programa que desenvolvemos para esse fim. Calculam-se, então, as *features*, utilizando-se *scripts* também desenvolvidos para serem utilizados neste trabalho. Os modelos de classificação foram gerados com o Weka, usando o algoritmo LMT, com os 46 marcadores iniciais e as cinco funções para medir riqueza de vocabulário para a classificação em gêneros, tipos textuais e necessidades personalizadas e com os 46 marcadores estatísticos e os 62 marcadores léxicos para a classificação em necessidades. O Leva-e-traz suporta html, docs, txt e pdf. Os modelos de classificação utilizados são os modelos para os quais obtivemos as melhores taxas de acerto (Capítulo 8).

No próximo capítulo apresentamos os resultados de uma avaliação com usuários sobre a compreensibilidade dos esquemas de classificação utilizados.

[51] O código do Leva-e-traz está disponível em http://www.nilc.icmc.usp.br/nilc/projects/recursoslinguarudo.html
[52] http://www.google.com/apis/

III

Avaliação

7. Utilidade teórica da abordagem segundo os usuários

"Searching is merely a means to an end – a way to satisfy an underlying goal that the user is trying to achieve." Rose & Levinson (2004)

Como mencionado no Capítulo 2, a avaliação de abordagens de RI tem um componente essencial de usabilidade. Se o usuário não percebe o que o sistema lhe oferece, a solução proposta neste trabalho não resolveria os seus problemas. Por isso, antes de propor a categorização em necessidades de busca, ou em gêneros, como solução para diminuir o esforço do usuário em encontrar documentos relevantes, resolvemos fazer um primeiro questionário para avaliar a compreensibilidade desses conceitos.

Nesse questionário tínhamos como objetivos verificar:

- a disposição dos usuários em investir um dia de trabalho na criação de esquemas de classificação personalizados;
- se o esquema de classificação segundo sete necessidades era claro para os usuários; se eles saberiam utilizá-lo e se gostariam que ele fosse diferente (com mais ou menos necessidades);
- se os usuários consideravam que classificar os resultados de busca segundo o esquema de sete necessidades seria útil para suas buscas;
- se esquemas de classificação de textos em gêneros também seriam úteis para buscas, se compreendiam todos os gêneros das taxonomias do Lácio-Ref (Seção 6.1.1), de Karlgren (Seção 5.2.3) e Stamatatos *et al* (2000b, Seção 5.2.5) e que gêneros consideravam particularmente úteis para suas buscas.

O questionário foi aplicado a alunos de graduação do primeiro semestre de um curso de sistemas de informação, segundo semestre de um curso de letras, último ano de um curso de medicina e a alunos de um curso de especialização em fotografia. O questionário é apresentado no Apêndice B. Setenta e três pessoas responderam ao questionário, porém dez dos questionários foram descartados, pois nesses 10 casos foram respondidas apenas duas ou três perguntas. Consideramos que esses usuários

não estavam muito empenhados em participar da pesquisa. Dos 63 usuários considerados, 18 eram alunos do curso de sistemas de informação, 25 do curso de letras, 14 do curso de medicina e 6 do curso de fotografia. Um resumo do perfil desses estudantes é mostrado na Tabela 9.

Tabela 9 – Perfil dos estudantes que responderam ao questionário sobre compreensão dos esquemas

Os estudantes são	
Homens	28
Mulheres	35
Têm entre	
18 e 29 anos	58
30 e 39 anos	3
mais de quarenta anos	2
Utilizam serviços de busca na Web	
raramente	4
ocasionalmente	16
freqüentemente	43
Consideram-se quanto à busca na Web, como	
inexperientes	3
razoavelmente experientes	49
muito experientes	11

No caso dos alunos de sistemas de informação e dos de letras, o questionário foi aplicado por seus professores em sala de aula; os estudantes tiveram cerca de 15 minutos para responder. No caso dos estudantes de medicina e de fotografia, o questionário foi entregue no final da aula por um de seus colegas e foi respondido em casa, posteriormente. Os alunos do primeiro grupo informaram terem dedicado entre 15 e 20 minutos para responder o questionário.

Dentre as 63 pessoas consultadas, 41 informaram encontrar problemas em suas buscas, porém, apenas 26 citaram exemplos. A lista dos problemas citados pelos estudantes é apresentada no Quadro 17; os problemas estão ordenados (ordem decrescente) pelo número de vezes em que foram mencionados.

Quadro 17 – Lista de problemas encontrados durante busca na Web citados pelos estudantes

(5) O resultado contém os termos utilizados na busca, mas não fala do assunto procurado.
(4) Retorna blogs
(4) Quando procuro por produtos para comprar, encontro páginas que os descrevem, mas não os vendem
(4) A página retornada não existe.

(3) Procuro por informações e recebo *links* que não têm informações, mas, sim, produtos à venda
(3) O texto apresenta informações superficiais sobre o assunto procurado.
(2) Retorna páginas desatualizadas.
(1) Retorna lixos virtuais
(1) Resultados que não são sobre o assunto procurado, mas que têm a palavra/palavras presentes uma única vez no texto. Por exemplo, quando procuro páginas sobre mitologia, encontro páginas em que a palavra aparece uma vez em alguma frase, mas que não são sobre mitologia.
(1) Quando utilizo palavras que são de mais de uma área, encontro também resultados que são da área que não me interessa
(1) Quando, procurando produtos para comprar, algumas vezes, encontro coisas demasiado abrangentes.
(1) Procurar por um assunto como "albergue da juventude" e encontrar todo tipo de coisas, inclusive hotéis.
(1) Procurar letras de música e encontrar páginas que vendem cds.
(1) Não consigo ver a relação entre o que procurei e os resultados mostrados

As 41 pessoas que disseram encontrar problemas em suas buscas informaram que estariam dispostas a utilizar um dia de trabalho para gerar esquemas de classificação personalizados. O que confirma o interesse de usuários por esquemas personalizados, ainda que tenham que criar um *corpus* com exemplos de páginas relevantes e irrelevantes para seu problema. Dessas, 7 deram exemplos de sistemas personalizados que gostariam de ter à sua disposição. Os exemplos são mostrados no Quadro 18, assim como aparecem nos questionários, ordenados pelo número de vezes em que foram mencionados.

Quadro 18 – Sistemas personalizados mencionados como de interesse

(2) Página com resumo de livros versus páginas com a biografia do autor
(1) Análise de obras literárias
(1) Informações técnicas versus informações para leigos
(1) Informática*
(1) Literatura brasileira, Literatura Francesa*
(1) Literatura dos anos 80*
(1) Medicina*
(1) Música, jogos eletrônicos e esporte*
(1) Poemas da língua portuguesa
(1) Psicologia*
(1) Textos técnicos e metodológicos da área de letras

Como pode ser visto no Quadro 18, seis dos 11 exemplos de sistemas personalizados podem estar relacionados à distinção entre assuntos, que não é nosso objeto de pesquisa, uma vez que lidamos como problemas personalizados relacionados a tipos de texto. Dizemos que podem estar porque não foram mais bem explicados e não tivemos como entrar em contato com esses usuários para esclarecer a dúvida. Ou seja, é possível que nossa proposta não tenha ficado clara através do exemplo dado no questionário, talvez por não termos dado um contra-exemplo, nem

termos dito que não faríamos distinção entre assuntos. Tentamos sanar essa dúvida — se é claro para os usuários para que serve o esquema personalizado — na avaliação que será descrita no Capítulo 9.

A grande maioria dos estudantes (61) disse que o esquema de classificação em necessidades facilitaria suas buscas na Internet. O número de estudantes que disseram que um dado esquema de classificação não seria útil para buscas na Internet é mostrado na Tabela 10. Quando perguntados se não entendiam o que era algum dos tipos de necessidade, apenas um dos tipos foi mencionado como não sendo claro: "uma página específica que o usuário quer visitar, mas não se lembra da URL". Essa questão foi levantada por seis pessoas. Porém, apesar de dizerem não terem dúvidas sobre o que os outros tipos de necessidades abrangiam, 14 dos estudantes não escolheram a opção correta de necessidade de busca para pelo menos uma de quatro das sete perguntas que fizemos: (i) "encontrar a página oficial do Palmeiras", 1 usuário escolheu o tipo de necessidade 6 (URL) e 2 escolheram o tipo de necessidade 3 (panorama), enquanto esperávamos que escolhessem o tipo 5 (pessoa/empresa/organização);
(ii) "encontrar uma crítica ou resenha sobre um cd ou show do Skank", 2 usuários escolheram o tipo 4 (notícia) enquanto esperávamos que escolhessem o tipo 3 (panorama);
(iii) "encontrar uma lista de hotéis em Araraquara", 2 usuários escolheram o tipo 7 (serviços) enquanto esperávamos que escolhessem o tipo 3 (panorama/apanhado);
(iv) "encontrar um site para envio de cartões virtuais" 1 usuário escolheu o tipo 3 "panorama" enquanto esperávamos que escolhessem o tipo 7 (serviços).

Todas as escolhas "erradas" mencionadas são compreensíveis, são escolhas lógicas. Um fato interessante é que, coincidentemente, os 11 usuários que se descreveram como muito experientes cometeram algum erro, os outros 3 que cometeram erros se consideram razoavelmente experientes. Os dados dão indícios de que, acrescentando-se uma descrição com exemplos de cada tipo de necessidade, os usuários saberiam escolher qual tipo de necessidade melhor se adequa à sua intenção de busca em um dado instante. Essa hipótese é investigada novamente na avaliação apresentada no Capítulo 10.

Tabela 10 - Número de estudantes que não consideram algum dos esquemas útil

Classificação segundo	Número de estudantes
As 7 Necessidades de busca (Seção 6.1.3)	2
O esquema de gêneros do Lácio-Ref apresentado através de exemplos de tipos de texto	3
Karlgren (Seção 5.2.3)	8
Gêneros do Lácio-Ref (Seção 6.1.1)	12
Stamatatos *et al* (2000b, Seção 5.2.5)	13

É interessante notar que o mesmo esquema de classificação, apresentado de formas diferentes, obteve diferentes avaliações. O esquema de gêneros do Lácio-Ref não foi considerado útil por 12 pessoas, porém, apenas três pessoas tiveram a mesma opinião quando o mesmo foi apresentado através de exemplos de tipos de textos comuns aos gêneros. Perguntamos também aos estudantes qual ou quais dos esquemas consideram mais útil (eis) e ou mais fácil (eis) de utilizar; a resposta pode ser vista na Tabela 11. É importante dizer que não foram dados nomes ou indicações de fontes sobre os esquemas de classificação no questionário. Não foi informado também o objetivo do questionário; foi dito apenas que esse questionário fazia parte de um trabalho de doutorado sobre estratégias de melhoria para a busca de textos na Web.

Tabela 11 - Número de estudantes que julgaram o esquema como mais fácil

Classificação segundo	Número de pessoas
As 7 Necessidades de busca (Seção 6.1.3)	25
O esquema de gêneros do Lácio-Ref apresentado através de exemplos de tipos de texto	29
Karlgren (Seção 5.2.3)	15
Gêneros do Lácio-Ref (Seção 6.1.1)	13
Stamatatos *et al* (2000b, Seção 5.2.5)	6

Considerando as três taxonomias de gêneros mostradas no questionário, quando perguntadas sobre quais dos gêneros não tinham significado totalmente claro, nove pessoas disseram não estar claro o que era o gênero "textos privados", oito o que era o gênero "jurídico", oito o "técnico-administrativo", sete o "*FAQs*", cinco o "de referência", cinco o "documentos oficiais", cinco o "textos públicos", três o "discursos planejados", três o "instrucional", três o "prosa acadêmica", duas o "discussões", duas o "outros textos", duas "páginas interativas e formulários" e uma o "listas e tabelas".

Foram sugeridos como novos gêneros/tipos textuais: “revista, classificada por temas”, “jornal, classificado por temas”, “imagens cirúrgicas”, “crônicas”, “cidade, industrialização, turismo” e “artigos científicos atualizados”. Em um conjunto tão pequeno de sugestões, vemos uma grande variedade de conceitos de gênero, pois temos gêneros relacionados a domínio como “imagens cirúrgicas”, meio e domínio como “revista, classificada por temas” e ainda impondo características que vão além do domínio, meio e propósito “artigos científicos atualizados”.

Parece-nos que a população consultada (jovens universitários), que, aliás, é relativamente representativa de uma classe importante da população brasileira na Web, votou favoravelmente ao uso de uma separação de resultados neste tipo de categorias, tanto em necessidades quanto em gêneros. Embora algumas das objeções e fraquezas do teste fossem patentes (e discutidas acima), os resultados deram-nos não só confiança para prosseguir com a abordagem descrita nos capítulos anteriores, mas também realimentação para um novo teste com usuários (descrito no Capítulo 10) já com o protótipo Leva-e-traz.

No próximo capítulo apresentamos a avaliação dos classificadores criados para cada esquema de classificação: em gêneros, em tipos textuais, em sete necessidades e em necessidades personalizadas. A avaliação é feita sob o ponto de vista do sistema, através da apresentação das medidas: taxa de acerto, precisão e revocação.

8. Taxa de acerto, precisão e revocação dos classificadores

"Cada problema que resolvi, tornou-se numa regra, que serviu depois para resolver outros problemas." René Descartes

Neste capítulo, apresentamos a avaliação dos classificadores, dados cada conjunto de esquema de classificação, algoritmo de classificação e marcadores de estilo utilizados. Os classificadores são comparados em termos de sua taxa de acerto (*accuracy*), mas apresentamos, também, precisão e revocação para que trabalhos de RI tenham como comparar seus resultados com os nossos, uma vez que essas são as duas medidas mais utilizadas na avaliação de sistemas de RI. Nas próximas seções, discutimos os resultados dos classificadores para a classificação em gêneros (Seção 8.1), tipos textuais (Seção 8.2), necessidades gerais de busca (Seção 8.3) e necessidades de busca personalizadas (Seção 8.4). Em todos os experimentos descritos neste capítulo, a avaliação foi feita utilizando a estratégia de validação cruzada estratificada em 10 partes (*ten-fold cross validation*).

A taxa de acerto é a proporção de acertos (verdadeiros positivos e verdadeiros negativos) dado um certo conjunto de casos de teste. Já a precisão, como descrita no Capítulo 2, indica a proporção de documentos relevantes recuperados e, por isso, considera apenas a proporção de verdadeiros positivos, dado um certo número de documentos retornados, e a revocação a proporção de verdadeiros positivos retornados, dado o número total de verdadeiro positivos.

8.1 Gêneros

Como explicado no Capítulo 6 (Seção 6.1.1), temos quatro versões do *corpus* de textos classificados quanto ao gênero: (1) a versão original do Lácio-Ref segundo gêneros; (2) uma versão com mais textos instrucionais; (3) uma versão em que substituímos os textos do gênero informativo originais por textos jornalísticos da Web; e (4) uma última versão em que fazemos as duas alterações mencionadas.

Em nossos primeiros experimentos com gêneros, feitos com a versão original do *corpus*, utilizamos os 46 marcadores estatísticos e os cinco marcadores de riqueza vocabular (a lista completa de marcadores é mostrada no Capítulo 6). Foram então dois conjuntos de *features*, um com 46 marcadores de estilo e outro com 51 marcadores de estilo (46 +5). Foram feitos experimentos com os algoritmos J48, SMO e LMT. Como pode ser visto na Tabela 12, nesses primeiros experimentos a melhor taxa de acerto foi obtida com 51 *features* e o algoritmo LMT, 97,21%. Por isso, em nossos experimentos com as outras três versões do *corpus*, optamos pelo conjunto de 51 *features*. Com a inclusão de mais gêneros do tipo Instrucional e a troca dos textos Informativos do *corpus* Lácio-Ref por textos jornalísticos da Web, a taxa de acerto foi menor, 94,87%, o que se deve a dois fatores: (1) o gênero instrucional já era problemático; com os experimentos com a versão original, os 3 textos instrucionais eram classificados erroneamente; e (2) trocamos 3.792 textos informativos por 150 textos da Web de diversas fontes (diversas seções de jornais, diversos jornais, de diferentes cidades e estados). Acreditamos ser esse o resultado que devemos considerar, por ser mais representativo da situação dos textos informativos presentes na Web. Foi esse o classificador que foi incluído no protótipo Leva-e-traz. Ainda assim, temos uma taxa de acerto alta, semelhante à dos melhores resultados de trabalhos de classificação em gênero do inglês (veja exemplos de resultados para o inglês no Capítulo 5).

Tabela 12 – Resultados da classificação em gêneros (as melhores taxas aparecem em negrito)

TAXA DE ACERTO			
	J48	**SMO**	**LMT**
Versão original do Lácio-Ref			
46 *features*	95,92%	95,72%	**96,29%**
51 *features*	95,86%	96,64%	**97,21%**
Versão do Lácio-Ref com mais textos do gênero Instrucional			
51 *features*	94,94%	**96,03%**	96,02%
Versão do Lácio-Ref com textos Informativos da Web			
51 *features*	94,24%	**96,48%**	95,36%
Versão do Lácio-Ref com mais Instrucionais e com Informativos da Web			
51 *features*	88,42%	93,35%	**94,87%**
PRECISÃO			
	J48	**SMO**	**LMT**
Versão original do Lácio-Ref			
46 *features*	**0,82**	0,81	0,72
51 *features*	0,80	0,83	**0,91**
Versão do Lácio-Ref com mais textos do gênero Instrucional			
51 *features*	0,76	**0,81**	**0,81**
Versão do Lácio-Ref com textos Informativos da Web			

51 *features*	0,93	**0,96**	**0,96**
Versão do Lácio-Ref com mais Instrucionais e com Informativos da Web			
51 *features*	0,88	0,93	**0,94**
REVOCAÇÃO			
	J48	**SMO**	**LMT**
Versão original do Lácio-Ref			
46 *features*	0,77	**0,85**	0,69
51 *features*	0,78	**0,86**	**0,86**
Versão do Lácio-Ref com mais textos do gênero Instrucional			
51 *features*	0,73	**0,85**	**0,85**
Versão do Lácio-Ref com textos Informativos da Web			
51 *features*	0,93	**0,96**	0,95
Versão do Lácio-Ref com mais Instrucionais e com Informativos da Web			
51 *features*	0,88	0,94	**0,95**

O algoritmo LMT utilizou 37 das 51 *features* como marcadores de estilo para a classificação em gêneros: número de caracteres no texto, estimativa do número de palavras diferentes, tamanho médio das palavras em caracteres, números de palavras com mais de 6 caracteres, de frases, médio de palavras por frase, de vezes em que as palavras "é" e "são" aparecem, de vezes em que a palavra "que" aparece, *amplifiers*, *conjuncts*, verbos persuasivos, verbos privados, verbos públicos, artigos definidos, artigos indefinidos, pronomes na segunda pessoa, pronomes na terceira pessoa, pronomes demonstrativos, alguns pronomes indefinidos e locuções, preposições, advérbios e locuções de lugar, e de advérbios e locuções de tempo; total de advérbios, números de interjeições, contrações, conjunções adverbiais causais, conjunções proporcionais, conjunções temporais, conjunções adverbiais concessivas, conjunções adverbiais condicionais, conjunções conformativas e conjunções comparativas e os cinco marcadores de riqueza vocabular apresentados no Capítulo 6.

8.2 Tipos Textuais

Nos experimentos com tipos textuais utilizamos duas versões do *corpus* do Lácio-Ref classificado por tipos textuais, a versão original do Lácio-Ref com 28 tipos textuais e 4.445 textos e a versão com 29 tipos textuais e 7.001 textos (ambos são descritos no Capítulo 6). Porém, utilizamos apenas os algoritmos J48 e SMO, pois o treinamento com algoritmo LMT, nesse caso, dadas as 29 classes, utilizando a estratégia de validação cruzada estratificada em dez partes, é muito lento. Em nossos experimentos com a versão original do *corpus*, utilizamos os mesmos dois conjuntos de *features* utilizados nos experimentos para a classificação em gêneros.

Tabela 13 - Resultados da classificação em tipos textuais (as melhores taxas aparecem em negrito)

TAXA DE ACERTO			
Versão original do Lácio-Ref			
	J48	**SMO**	**LMT**
46 *features*	74,09%	**75,53%**	--
51 *features*	74,14%	**77,49%**	--
Versão do Lácio-Ref com mais textos extraídos da Web			
51 *features*	74,20%	73,08%	**83,95%**
PRECISÃO			
Versão original do Lácio-Ref			
	J48	**SMO**	**LMT**
46 *features*	**0,66**	0,56	--
51 *features*	**0,66**	0,55	--
Versão do Lácio-Ref com mais textos extraídos da Web			
51 *features*	0,58	0,50	**0,87**
REVOCAÇÃO			
Versão original do Lácio-Ref			
	J48	**SMO**	**LMT**
46 *features*	0,68	**0,91**	--
51 *features*	0,67	**0,92**	--
Versão do Lácio-Ref com mais textos extraídos da Web			
51 *features*	0,59	**0,83**	0,81

Para a classificação em tipos textuais, os melhores resultados também foram com o conjunto de 51 *features*, 77,49% (veja Tabela 13); por isso, nos experimentos com mais textos, esse foi o conjunto de *features* utilizado. A taxa de acerto foi consideravelmente menor na classificação em tipos textuais do que em gêneros, pois são 24 classes a mais do que a classificação em gêneros. Além disso, o *corpus* não é balanceado como pode ser visto no Capítulo 6. Existem tipos representados por 3 unidades e textos representados por mais de mil exemplos. Foi devido ao *corpus* não ser balanceado que construímos a nova versão com mais textos, porém o mesmo continuou não balanceado. O tipo com menos exemplos tem 3 textos e o com mais 2171 textos. Mesmo com os esforços para incluir mais textos ao *corpus*, a taxa de acerto não melhorou com o novo *corpus*, sem aumento estatisticamente significante para o algoritmo J48 e menor com o algoritmo SMO, 73,08%. Acreditamos que, também nesse caso, os resultados não foram melhores com a inclusão de mais textos, por não termos incluído um número suficientemente grande de exemplos das variações mais comuns de cada tipo de texto presente na Web. Mas, ainda assim, incluímos no *corpus* um conjunto de textos muito menos homogêneo quanto à fonte do que os originalmente presentes no *corpus*. Porém, quando repetimos o experimento com o algoritmo LMT, a taxa de acerto subiu para 83,95%. Ainda abaixo dos

resultados para gênero, mas melhor do que a de trabalhos semelhantes para inglês quando se lida com esquemas de classificação mais específicos quanto ao tipo de texto. Karlgren (2000), por exemplo, obteve uma taxa de acerto de 51,6% para a classificação em 15 categorias, utilizando o *corpus* Brown. O LMT utilizou como marcadores de estilo para esta tarefa todas as 51 *features*.

8.3 Necessidades de busca

Como explicado no Capítulo 6, a primeira versão do *corpus* de necessidades foi composta por 511 textos, e não foi considerado o fato de que um texto poderia atender a mais de uma necessidade. Os experimentos com esse *corpus* foram os primeiros experimentos realizados neste projeto, quando havíamos selecionado apenas o conjunto de 46 *features*. Os resultados com os algoritmos J48, SMO e LMT, utilizando essa primeira versão do *corpus*, suas 7 categorias (6 necessidades e a categoria "outros") podem ser vistos na Tabela 14. Fizemos também experimentos com esse mesmo *corpus* para 5 e 3 categorias; o primeiro caso une os tipos de necessidades 1, 2 e 3 (definições, instruções e panorama), o segundo une os tipos 1, 2, 3, 4 e 5 (definições, instruções, panorama, notícias e informações sobre pessoas e empresas) em uma única categoria. A classificação em 3 categorias corresponde às categorias informacional e transacional explicadas no Capítulo 5. Os resultados para esses esquemas também são mostrados na Tabela 14.

Tabela 14 – Resultados da classificação em necessidades utilizando o *corpus* de 511 textos (as melhores taxas aparecem em negrito)

TAXA DE ACERTO			
	J48	**SMO**	**LMT**
Considerando as seis necessidades de busca			
	45,32%	**57,78%**	57,23%
Informação, Notícias, Informações sobre pessoa ou empresa e Serviço			
	56,56%	--	--
Informação versus serviço			
	76,97%	--	--
PRECISÃO			
	J48	**SMO**	**LMT**
Considerando as seis necessidades de busca			
	0,39	**0,55**	0,54
Informação, Notícias, Informações sobre pessoa ou empresa e Serviço			
	0,68	--	--
Informação versus serviço			
	0,86	--	--
REVOCAÇÃO			

	J48	SMO	LMT
Considerando as seis necessidades de busca			
	0,45	**0,53**	0,46
Informação, Notícias, Informações sobre pessoa ou empresa e Serviço			
	0,7	--	--
Informação versus serviço			
	0,88	--	--

Na Web existem muitos mais tipos de texto do que os cobertos por nossa taxonomia de sete necessidades de busca, por isso, incluímos textos classificados como "outros". Exemplos de textos incluídos na classe "outros" são páginas de piadas (como descrito na Seção 6.1.3). Mesmo com um conjunto pequeno de textos classificados como "outros", essa inclusão fez com que a taxa de acerto caísse consideravelmente; por exemplo, sem a classe "outros" a taxa de acerto para a classificação em páginas que atendem a necessidades informacionais versus páginas que atendem a necessidades transacionais seria de 90,93% e não de 76,97% apenas. Nossa análise desses primeiros resultados nos fez acreditar que um dos problemas que poderiam estar causando a baixa taxa de acerto era o fato de que não considerávamos que uma página pudesse atender a mais de uma necessidade. Por isso o *corpus* foi reclassificado como descrito na Seção 6.1.3, dando origem à versão atual com 1.703 páginas.

Nos experimentos com a versão aumentada do *corpus*, ao invés de gerarmos um único classificador para todos os tipos de necessidades, geramos um classificador para cada necessidade e mantivemos os textos do tipo "outros" em todos os experimentos. Isso foi feito porque não nos interessava ter classificadores para os tipos combinados[53], por exemplo, 1-2. Assim, todos os textos que atendessem a um tipo seriam reunidos para o treinamento; por exemplo, os textos marcados como 1-2 aparecem no conjunto de treinamento do Tipo 1 como Tipo 1, e do Tipo 2 como Tipo 2. As taxas de acerto são mostradas na Tabela 15, a precisão na Tabela 16 e a revocação na Tabela 17. Nos experimentos com a versão atual do *corpus*, utilizamos os 46 marcadores estatísticos, os 5 marcadores de riqueza vocabular e os 62 marcadores léxicos. Como explicado na Seção 6.3, os marcadores léxicos foram selecionados a partir das palavras mais freqüentes do *corpus*. Foram utilizados 3 conjuntos de *features*: os dois primeiros, já descritos anteriormente, o de 46 *features* e

[53] Assumindo-se que o usuário escolheria um tipo de necessidade por vez.

o de 51 *features*; e o último de 108 *features* composto pelos 46 marcadores estatísticos e os 62 marcadores léxicos. Nos experimentos com o conjunto de 108 *features*, utilizamos os algoritmos J48, SMO e LMT; nos demais apenas os algoritmos J48 e SMO.

Tabela 15 - Taxa de acerto da classificação por necessidades (as melhores taxas aparecem em negrito)

	J48	**SMO**	**LMT**
Tipo 1 – Definições ou explicações sobre como ou porque algo acontece			
46 *features*	59,80%	**65,75%**	--
51 *features*	60,07%	**65,60%**	--
108 *features*	59,87%	**67,25%**	67,23
Tipo 2 – Instruções sobre como fazer algo ou como algo é feito			
46 *features*	70,62%	**75,97**%	--
51 *features*	71,96%	**76,50**%	--
108 *features*	76,53%	82,95%	**82,97**%
Tipo 3 – Panorama			
46 *features*	55,46%	**58,70**%	--
51 *features*	56,30%	**59,85**%	--
108 *features*	59,41%	**63,76**%	63,01%
Tipo 4 – Notícias			
46 *features*	77,55%	**79,07**%	--
51 *features*	76,67%	**79,79**%	--
108 *features*	81,22%	85,60%	**86,14**%
Tipo 5 – Informações sobre pessoas/empresas/organizações/instituições			
46 *features*	77,81%	**82,03**%	--
51 *features*	77,94%	**82,03**%	--
108 *features*	78,64%	84,69%	**84,97**%
Tipo 7 – Serviços			
46 *features*	81,38%	**85,42**%	--
51 *features*	81,78%	**85,49**%	--
108 *features*	88,57%	91,02%	**91,19**%

As melhores taxas de acerto foram obtidas com o conjunto de 108 *features*. A taxa de acerto média é de 79,38% para as seis necessidades, uma taxa de acerto consideravelmente melhor do que a taxa de acerto obtida com o *corpus* original. A pior taxa de acerto é para o tipo 3, páginas que forneçam panoramas, 63,76%. Isso ocorre porque páginas que forneçam panoramas fornecem dentre outras informações, informações dos tipos 1, 2, 4 e 5; o que se complica no nosso caso em particular, uma vez que definimos que panoramas poderiam ser de qualquer tamanho e, mesmo que tenham mais informações de um dado tipo, continuam sendo classificados como panoramas. O outro problema encontrado em nossa análise foi a classificação de biografias, que quando continham várias informações foram classificadas no *corpus* como panorama e como informações sobre pessoas. A segunda pior taxa de acerto foi

para o Tipo 1, 67, 25%. Nesse caso, a possível causa é também o fato de não levarmos em conta qual tipo de informação prevalece em um texto. As demais taxas de acerto ficaram acima dos 80%, sendo que a classificação de páginas como oferecendo ou não serviços online (transacional versus informacional) foi de 91,19%. Pelos resultados, imagina-se que uma taxonomia, com melhores resultados e ainda útil, teria as seguintes quatro categorias: definições e instruções, notícias, informações sobre pessoas ou empresas, e serviços.

Tabela 16 – Precisão da classificação por necessidades (as melhores taxas aparecem em negrito)

	J48	**SMO**	**LMT**
Tipo 1 – Definições ou explicações sobre como ou porque algo acontece			
46 *features*	0,51	**0,64**	--
51 *features*	0,52	**0,64**	--
108 *features*	0,51	0,65	**0,66**
Tipo 2 – Instruções sobre como fazer algo ou como algo é feito			
46 *features*	0,55	**0,78**	--
51 *features*	0,58	**0,78**	--
108 *features*	0,64	**0,82**	**0,82**
Tipo 3 – Panorama			
46 *features*	0,56	**0,60**	--
51 *features*	0,56	**0,60**	--
108 *features*	0,60	**0,63**	**0,63**
Tipo 4 – Notícias			
46 *features*	0,61	**0,77**	--
51 *features*	0,59	**0,78**	--
108 *features*	0,67	0,81	**0,84**
Tipo 5 – Informações sobre pessoas/empresas/organizações/instituições			
46 *features*	0,35	0	--
51 *features*	0,36	0	--
108 *features*	0,4	0,71	**0,75**
Tipo 7 – Serviços			
46 *features*	0,6	**0,72**	--
51 *features*	0,61	**0,72**	--
108 *features*	0,76	0,83	**0,89**

No protótipo Leva-e-traz, incluímos os modelos gerados a partir do conjunto de 108 *features* com o algoritmo LMT, apesar desse ter uma taxa de acerto muito próxima e, em alguns casos, pouco inferior ao SMO. O motivo da escolha foi devido ao fato de o *Kappa*[54] ser maior para o LMT do que para o SMO nesses experimentos. Foram utilizados pelo LMT 104 das *features* como marcadores; as únicas *features* não utilizadas foram: número de caracteres no texto, tamanho médio das palavras em caracteres, número de frases e número de vezes que a palavra "se" aparece.

[54] Taxa de concordância entre os 10 classificadores gerados em cada uma das fases do teste, utilizando validação cruzada estratificada em 10 partes.

Tabela 17 – Revocação da classificação por necessidades (as melhores taxas aparecem em negrito)

	J48	**SMO**	**LMT**
Tipo 1 – Definições ou explicações sobre como ou porque algo acontece			
46 *features*	**0,49**	0,38	--
51 *features*	**0,5**	0,38	--
108 *features*	0,51	0,44	**0,65**
Tipo 2 – Instruções sobre como fazer algo ou como algo é feito			
46 *features*	**0,47**	0,35	--
51 *features*	**0,48**	0,37	--
108 *features*	0,62	0,6	**0,78**
Tipo 3 – Panorama			
46 *features*	**0,62**	0,56	--
51 *features*	**0,62**	0,61	--
108 *features*	0,59	**0,66**	0,63
Tipo 4 – Notícias			
46 *features*	**0,54**	0,35	--
51 *features*	**0,54**	0,38	--
108 *features*	0,65	0,64	**0,81**
Tipo 5 – Informações sobre pessoas/empresas/organizações/instituições			
46 *features*	0,28	0	--
51 *features*	0,29	0	--
108 *features*	0,38	0,26	**0,67**
Tipo 7 – Serviços			
46 *features*	0,6	**0,62**	--
51 *features*	0,6	**0,61**	--
108 *features*	0,75	0,77	**0,87**

Como pode ser visto nas tabelas 15, 16 e 17, a precisão e a revocação são, na maioria dos resultados, menores que a taxa de acerto. Isso, em geral, acontece nos casos em que o número de exemplos positivos é raro e, em nosso *corpus*, o número de exemplos positivos é consideravelmente menor do que o número de exemplos negativos. Pois, apesar do *corpus* ser balanceado considerando-se as possíveis combinações de necessidades (veja Capítulo 6), quando se unem as diversas combinações para gerar um classificador para um dado tipo de necessidade, o *corpus* deixa de ser balanceado. Por exemplo, para treinar um classificador para identificar páginas que prestam serviços online, teríamos 395 exemplos positivos contra 1308 negativos.

Após termos concluído as avaliações mostradas nesse capítulo e nos capítulos 9 e 10, fizemos experimentos iniciais com dois novos conjuntos de marcadores, o conjunto de 15 marcadores sintáticos e o de 27 marcadores de aparência gráfica. Na Tabela 18 mostramos as taxas de acerto acrescentado-se os marcadores de aparência

gráfica aos conjuntos de 46 *features* (total de 73 *features*) e de 108 *features* anteriores (total de 135 *features*).

Tabela 18 – Resultados da classificação em necessidades, utilizando-se marcadores de aparência gráfica (as melhores taxas aparecem em negrito)

	J48	**SMO**	**LMT**
Tipo 1 – Definições ou explicações sobre como ou porque algo acontece			
73 *features*	58,99%	**65,8**%	65,46%
135 *features*	59,14%	**66,95**%	66,68%
Tipo 2 – Instruções sobre como fazer algo ou como algo é feito			
73 *features*	70,22%	76,24%	**78,04**%
135 *features*	75,8%	**83,18**%	82,89%
Tipo 3 – Panorama			
73 *features*	55,63%	59,24%	**59,65**%
135 *features*	57,58%	**63,79**%	63,16%
Tipo 4 – Notícias			
73 *features*	77,35%	**80,21**%	81,2%
135 *features*	81,06%	86,23%	**86,93**%
Tipo 5 – Informações sobre pessoas/empresas/organizações/instituições			
73 *features*	77%	82,66%	**82,95**%
135 *features*	78,52%	83,39%	**84,92**%
Tipo 7 – Serviços			
73 *features*	80,34%	**85,31**%	**85,31**%
135 *features*	88,36%	91,4%	**91,47**%

Na Tabela 19 mostramos as taxas de acerto acrescentado-se os marcadores sintáticos aos conjuntos de 46 *features* (total de 61 *features*) e de 108 *features* anteriores (total de 123 *features*).

Tabela 19 - Resultados da classificação em necessidades, utilizando-se marcadores sintáticos (as melhores taxas aprecem em negrito)

	J48	**SMO**	**LMT**
Tipo 1 – Definições ou explicações sobre como ou porque algo acontece			
61 *features*	59,28%	66,29%	**66,44**%
123 *features*	59,34%	67,51%	**67,53**%
Tipo 2 – Instruções sobre como fazer algo ou como algo é feito			
61 *features*	72,67%	80,63%	**80,89**%
123 *features*	76,79%	83,96%	**83,99**%
Tipo 3 – Panorama			
61 *features*	57,58%	**63,79**%	63,16%
123 *features*	59,73%	65,81%	**65,85**%
Tipo 4 – Notícias			
61 *features*	77,03%	82,01%	**83,46**%
123 *features*	81,86%	87,34%	**87,82**%
Tipo 5 – Informações sobre pessoas/empresas/organizações/instituições			
61 *features*	78,67%	82%	**84,52**%
123 *features*	78,38%	**85,94**%	85,47%
Tipo 7 – Serviços			
61 *features*	80,97%	**85,54**%	85,33%
123 *features*	87,78%	91,02%	**91,28**%

Como pode ser visto nas tabelas 18 e 19, nem os marcadores de aparência gráfica, nem os marcadores sintáticos resultaram em melhorias significativas para a taxa de acerto da classificação em necessidades. No caso dos marcadores de aparência gráfica, isso pode estar relacionado ao fato de as tags html, utilizadas como marcadores, não serem utilizadas de forma padronizada nas páginas, o que precisa ser investigado em trabalhos futuros, analisando-se grandes conjuntos de páginas html para selecionarmos as tags html mais freqüentes. No segundo caso, seria interessante rodarmos experimentos apenas com as categorias sintáticas para que pudéssemos comparar os resultados com os dos outros conjuntos de marcadores individualmente, visto que muitos dos 46 marcadores iniciais foram pensados para refletir exatamente características sintáticas. Ou seja, tanto as experiências com os marcadores sintáticos, quanto as experiências com os marcadores de aparência gráfica são preliminares, e a interpretação e validação dos resultados terão de ser feitas em trabalhos futuros.

8.4 Necessidades personalizadas

Em nossos primeiros testes com necessidades personalizadas utilizamos os *corpora* de direito e de páginas com descrição de produtos à venda, descritos no Capítulo 6. Os experimentos foram realizados, utilizando o conjunto de 46 *features* e os algoritmos J48, SMO e LMT.

Tabela 20 – Resultados para a classificação em necessidades personalizadas (as melhores taxas aparecem em negrito)

	J48	**SMO**	**LMT**
***Corpus* de direito**			
Taxa de acerto	76,35%	82,85%	**83,20%**
Precisão	0,77	0,83	**0,84**
Revocação	0,76	**0,84**	**0,84**
***Corpus* sobre páginas com descrição de produtos a venda ou não**			
Taxa de acerto	88,28%	83,7%	**89,91%**
Precisão	0,87	**0,90**	**0,90**
Revocação	0,85	0,69	**0,86**

Para ambos os problemas, obtivemos uma taxa de acerto boa, acima de 83%. Utilizando o arquivo de treinamento de Martins & Moreira (2004) com suas 11.205 *features*, as taxas de acerto com os algoritmos J48 e SMO foram, respectivamente, 96,26% e 94,19%.

8.5 Considerações sobre os resultados

Sabe-se do significado das medidas taxa de acerto, precisão e revocação para a comparação de sistemas, mas não é claro o que as mesmas significam para o usuário. No caso do problema específico com o qual tratamos que é poupar o usuário de lidar com documentos irrelevantes, é natural que se dê um maior peso à precisão do que para a revocação. Particularmente, no caso do protótipo Leva-e-traz, em que os resultados são classificados e ordenados, mas os resultados que atendem a outros tipos de necessidades não são eliminados, privilegiar a precisão e não a revocação não é prejudicial. Pois, ao não encontrar respostas, o usuário pode ver os resultados classificados de outros tipos, como faria em um sistema tradicional visitando *links* que a princípio não imaginava serem relevantes após não ter encontrado todas as respostas nos que julgou inicialmente serem relevantes. Porém, a precisão utiliza apenas os casos em que o sistema disse que um texto atendia a uma necessidade e ele realmente a atendia (verdadeiros positivos), não considera os casos em que o sistema disse que o texto não atendia a uma necessidade e ele não atendia (verdadeiros negativos); por isso, parece-nos ainda mais natural que se utilize a taxa de acerto, que faz uso de ambos os casos.

Ainda assim, permanece não respondida a questão de o que significa um determinado valor de taxa de acerto para o usuário, qual o valor mínimo para que a classificação seja útil para o usuário. Ou seja, não sabemos se, dada uma taxa de acerto média de 79,38%, a classificação em necessidades é útil para o usuário. Por isso, fez-se necessário conduzir outras avaliações, desta vez centradas no usuário e não no sistema, que serão mostradas nos Capítulos 9 e 10.

No próximo capítulo, são apresentados resultados da avaliação da opção de classificação personalizada com usuários.

9. Resultados com a busca personalizada

"Isolar a matemática das questões práticas é um convite à esterilidade de uma vaca atirada para longe dos bois." Pafnuty Chebyshev

A avaliação da busca personalizada, sem utilizar-se de *corpora* criados por usuários reais, não seria uma avaliação completa, dado que: (i) conhecemos bem a abordagem seguida e poderíamos, ainda que inconscientemente, selecionar problemas e textos mais fáceis de tratar; (ii) os usuários podem não entender o que a opção de classificação personalizada oferece, como nos seis exemplos de sistemas personalizados relacionados a assunto e não a tipo de texto, citados por estudantes em nossa avaliação prévia mostrada no Capítulo 7; (iii) ou podem ter dificuldades para criar o *corpus*; e (iv) podem, ainda, mudar sua disposição quanto a criar *corpus* de treinamento, após terem tido essa experiência pela primeira vez.

Na avaliação de busca personalizada, contamos com a colaboração de seis usuários, cinco portugueses e um brasileiro, dos quais dois têm formação em letras e quatro em computação. Solicitamos a cada um por *e-mail* que descrevessem o problema que seria tratado por seus *corpora* e que criassem cada um, um *corpus* com 200 textos, sendo 100 exemplos positivos e 100 negativos. Após a criação do *corpus*: (i) perguntamos aos usuários quanto tempo levaram para criar o *corpus* e de qual (quais) variante(s) da língua eram os textos; (ii) criamos o classificador para cada *corpus* e enviamos um *e-mail* aos usuários agradecendo por sua participação, informando a taxa de acerto para o seu problema e perguntando se, dada a experiência que tiveram, criariam novos *corpora* para outros problemas no futuro.

No pedido enviado, descrevíamos de forma simples a busca personalizada, com exemplos de problemas para os quais ela funcionaria e um exemplo de problema para o qual não funcionaria. Veja os exemplos dados no Quadro 19.

Quadro 19 – Exemplos de problemas fornecidos aos usuários que criaram os corpora

Tratamos apenas de problemas de duas classes, que estejam relacionados a tipos de texto. Por exemplo, no caso de um médico, distinguir entre textos que falam sobre medicina que são técnicos e textos voltados para o público comum. Não tratamos de problemas relacionados a assunto; por exemplo, distinguir entre textos de cardiologia e textos que falam sobre problemas de coração, mas não são de cardiologia. A distinção deve estar relacionada a tipos textuais, páginas mais formais/informais, que emitam opinião ou apenas descrevam fatos, etc. Exemplos de necessidades personalizadas: • Sou advogado, constantemente procuro por material de apoio para o meu trabalho, e na lista de resultados da busca estão sempre presentes textos sobre os assuntos procurados, mas que são voltados para leigos. • Faço compras pela Internet constantemente e sempre me irrito, pois nas consultas aparecem, freqüentemente, blogs com o relato de alguém que comprou o produto, propagandas, notícias que falam de seus lançamentos e não o que quero, que são lojas que vendam o produto online. • Meu hobby é ir às touradas sempre que posso, uso a Web para encontrar relatos sobre a última tourada em que não pude ir, ou dados sobre as que irão acontecer. Não gostaria de ter na minha lista de resultados textos sobre a história das touradas, sobre as normas seguidas etc., pois já conheço muito sobre o assunto.

9.1 Os corpora

Foram criados sete *corpora* em resposta à solicitação por *e-mail*.[55] A descrição dos problemas tratados por cada classificador personalizado é apresentada no Quadro 20. Um dos usuários formados em letras criou dois *corpora* para problemas distintos. Não houve interferência nossa em nenhum momento da criação dos *corpora*. Foi pedido apenas aos usuários que confirmassem se iriam ou não participar e informassem o problema que iriam tratar. Alguns dos usuários, nesse momento de confirmação, questionaram-nos, perguntando se o problema era o que desejávamos, respondemos apenas que sim.

Quadro 20 – Descrição das sete necessidades personalizadas tratadas

Problema 1. Obter textos teóricos em html sobre filosofia da linguagem e sobre os principais pensadores e não textos (também em html) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico. **Problema 2.** Obter textos teóricos em html sobre língua portuguesa e sobre os principais pensadores e não textos (também em html) que apresentem programas de cursos, colóquios, conferências, livros, etc. sobre este tópico. **Problema 3.** Diferenciar textos que apresentem fatos sobre Fado de textos que emitam opiniões. No primeiro caso estão textos contendo informação histórica, biografias, noticias, etc. No segundo, entrevistas, críticas a discos e espetáculos, etc.

[55] Os *corpora* estão disponíveis em http://www.linguateca.pt/Repositorio/YesUser/

Problema 4. Encontrar textos que sejam uma descrição sobre determinado tema de História Geral. Entretanto, páginas com eventos, conferências, catálogos de livros não interessam, bem como informações sobre cursos de História, *link*s para páginas de História ou ementas de disciplinas. Além disso, relatos de pessoas sobre seu gosto pela História também não são de interesse.

Problema 5. Distinguir glossários, receitas e técnicas sobre culinária japonesa de anúncios de livros, informação nutricional, criticas a restaurantes, páginas de restaurantes, informação sobre alimentação, cursos, festivais gastronômicos ou culturais sobre o mesmo tema.

Problema 6. Texto que interessam são história/fatos sobre surrealismo, como "Salvador Dali e o Surrealismo", "Manifesto do Surrealismo" e "Enciclopédia Universal Multimídia On-line". Blogs, exposições ou opiniões como "BdE - Blogue de Esquerda (II) 80 ANOS DE SURREALISMO", "Adelto Gonçalves,- comemorações", "A estranha sombra do surrealismo português, não interessam".

Problema 7. Documentos relevantes são: 1 - Documentos que explicam os princípios físicos que permitem que os aviões vôem; 2 - Explicações técnicas de partes de componentes de aviões, tais como altímetros, tipos de motores ou rotores de helicópteros, etc.; 3 - História da aviação - biografia de pioneiros da aviação, os avanços aeronáuticos ao longo do tempo; 4 - História dos aviões - História de certos aviões importantes para a historia, as suas características, o motivo do seu desenvolvimento, o seu impacto na história da aviação.

Documentos não-relevantes são: 1 - Notícias relacionadas com compras de aviões e empresas de aviação comercial; 2 - Notícias e descrições detalhadas de acidentes aéreos; 3 - Relatos de desvio de aviões e terrorismo aéreo; 4 - Opiniões sobre pilotagem, histórias e relatos de clubes de aviação, diversos documentos sobre psicologia do avião, deveres dos pilotos, etc.

Na tabela 21, descrevemos detalhes da criação de cada um dos *corpora*, mostrando o número de exemplos positivos e negativos, o número de horas gasto para fazer os *corpora* e se há uma ou mais variantes do português.

Tabela 21 – Descrição dos *corpora* criados por usuários

Corpus	Nº de exemplos positivos	Nº de exemplos negativos	Nº de horas gastos na criação do corpus	Variantes do português
***Problema* 1**	55	49	3 horas	brasileira e portuguesa
***Problema* 2**	60	101	5 horas	brasileira e portuguesa
***Problema* 3**	100	100	+- 6 horas	grande maioria da portuguesa
***Problema* 4**	105	100	6 horas	brasileira
***Problema* 5**	100	100	+- 6 horas	grande maioria da brasileira
***Problema* 6**	76	95	entre 4 e 5 horas	maioria da brasileira
***Problema* 7**	83	81	+- 10 horas[56]	maioria da brasileira

9.2 Resultados

Os classificadores foram criados utilizando as 46 *features* iniciais mais as 5 *features* de riqueza vocabular e o algoritmo SMO. O algoritmo SMO foi escolhido porque é

[56] Os usuários que criaram os *corpora* 6 e 7 nos informaram terem descoberto serem poucas as páginas na Web de exemplos positivos sobre o problema em que estavam interessados. Porém, apenas o usuário que criou o *corpus* 7 insistiu na procura.

muito mais rápido que o LMT no treinamento. Como pode ser visto na Tabela 22, para quatro dos problemas tratados, a taxa de acerto foi de 85% ou mais, para dois casos, ficou entre 70% e 80% e em um caso foi de apenas 65,5%.

Tabela 22 – Resultados da classificação personalizada com *corpus* de usuários

	Taxa de acerto	
***Problema* 1**	**87,38%**	
	Precisão	**Revocação**
Página desejada	0,89	0,87
Outro tipo de página	0,86	0,89
***Problema* 2**	**85%**	
	Precisão	**Revocação**
Página desejada	0,80	0,80
Outro tipo de página	0,88	0,88
***Problema* 3**	**79,90%**	
	Precisão	**Revocação**
Página desejada	0,78	0,83
Outro tipo de página	0,82	0,77
***Problema* 4**	**85,37%**	
	Precisão	**Revocação**
Página desejada	0,89	0,82
Outro tipo de página	0,82	0,89
***Problema* 5**	**87,50%**	
	Precisão	**Revocação**
Página desejada	0,84	0,92
Outro tipo de página	0,91	0,83
***Problema* 6**	**65,50%**	
	Precisão	**Revocação**
Página desejada	0,62	0,59
Outro tipo de página	0,68	0,71
***Problema* 7**	**73%**	
	Precisão	**Revocação**
Página desejada	0,72	0,77
Outro tipo de página	0,75	0,69

No caso do Problema 3, para o qual a taxa de acerto foi de 79,9%, o *corpus* foi o único composto em sua maioria por textos escritos em português de Portugal; porém, no momento da escolha dos marcadores estilísticos, tínhamos levado em consideração apenas a variante brasileira da língua. No caso dos problemas 6 e 7, taxa de acertos de 65,5% e 73%, respectivamente, os usuários informaram ter tido dificuldades em encontrar exemplos positivos. Analisando ambos os *corpora*, verificamos que os exemplos são muito diferentes estilisticamente uns dos outros. A taxa de acerto baixa nesses três casos indica que é necessário estudar mais a busca personalizada, tanto os marcadores de estilo levando em consideração outras variantes

da língua, quanto mais temas, para que se verifique qual o tamanho mínimo de *corpus* necessário e se existem variações dessa indicação.

Após os experimentos, informamos por e *e-mail* a precisão para o problema tratado e perguntamos a cada usuário, individualmente: "Após ter participado deste experimento criando o *corpus*, se tivesse disponível uma ferramenta para criar formas de classificação personalizadas para buscas específicas como esta que faça com freqüência na Web, você a utilizaria?". Todos os usuários responderam sim a essa pergunta. Dois deles teceram os seguintes comentários:

- Sim, principalmente quando o objetivo final fosse fazer um trabalho de investigação em que precisaria selecionar bem as páginas mais adequadas;
- Com certeza que usaria, apesar de considerar a tarefa de criar o *corpus* trabalhosa e cansativa, quando feita de uma única vez.

No próximo capítulo são mostrados os resultados da avaliação com usuários do uso da classificação em necessidades de busca na busca na Web em português.

10. Estimativa do esforço de busca dos usuários

"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

Na avaliação mostrada no Capítulo 7, apenas duas de 63 pessoas disseram não considerar a classificação em necessidades útil. Porém, 14 pessoas escolheram o tipo de necessidade incorreto para, pelo menos, uma de quatro das sete consultas questionadas. Neste capítulo, apresentaremos uma avaliação realizada com o auxílio do protótipo Leva-e-traz e um questionário que é apresentado no Apêndice C. O objetivo da avaliação foi verificar: (i) se os usuários compreenderam com clareza a taxonomia de sete necessidades; (ii) se o usuário visitaria menos resultados irrelevantes do que em um sistema sem a classificação em necessidades de busca no caso dos resultados serem classificados com ela; e (iii) se após utilizar o protótipo, os usuários considerariam o esquema de necessidades útil.

Dez pessoas participaram da avaliação, sendo três com formação em letras, seis com formação em computação e um estudante de graduação do último ano de ciência da computação[57]. Três dessas são mulheres e 7 são homens; sete têm entre 20 e 29 anos e três, entre 30 e 40; todos disseram utilizar máquinas de busca freqüentemente; 9 se consideraram razoavelmente experientes e 1 se classificou como muito experiente.

Dado o fato de a taxonomia ter sido proposta para a busca na Web, avaliações com usuários deveriam ser feitas com uma amostra de tamanho significante de usuários com perfis diversificados. Porém, contava-se neste projeto apenas com voluntários, por isso o número reduzido de usuários. Apesar disso, realizou-se a avaliação, para que se levantassem indicações para uma proposta de avaliação que pudesse ser utilizada em futuros trabalhos nos quais haja patrocínio para a avaliação.

[57] Nenhum dos usuários participou em algum outro instante deste projeto, seja direta ou indiretamente através da participação em outras avaliações.

Por exemplo, uma máquina de busca que solicitasse a usuários participarem de uma avaliação em troca da participação em um sorteio de um determinado prêmio.

10.1 Estrutura da avaliação

Foi solicitado a cada usuário que encontrasse respostas para seis tópicos, mostrados no Quadro 21. Os usuários deveriam: (i) efetuar as três primeiras pesquisas utilizando a tela principal do Leva-e-traz, que não classifica os resultados, apenas reúne os 10 primeiros resultados do Google e os 10 primeiros do Alltheweb[58]; (ii) ler a página de *Help* do Leva-e-traz sobre as sete necessidades (veja Figura 13 no Apêndice A), depois de concluídas as primeiras pesquisas, e (iii) só então realizar as três últimas pesquisas, desta vez utilizando a *tab* necessidades do Leva-e-traz que disponibiliza a classificação em sete necessidades (veja Figura 8 no Apêndice A). O número de tópicos ficou restrito a seis, para evitar que usuários se cansassem durante a avaliação e a realizassem com pouco zelo a partir de um dado momento.

Para se comparar o esforço do usuário para julgar a relevância, dividimos os 10 usuários e os seis tópicos em dois grupos, para que assim cada um dos dois grupos de usuários respondesse a perguntas diferentes em cada uma das duas fases mencionadas acima. A razão desse cenário foi evitar que a segunda busca fosse realizada com mais facilidade, isto é, que após um usuário ter buscado por respostas para um tópico com um dos métodos, ele pudesse estar condicionado e assim efetuar com mais facilidade a busca pelo tópico em um segundo momento. Ou seja, a divisão dos usuários em dois grupos permitiu que um usuário A respondesse os tópicos 1 a 3 sem classificação, enquanto o usuário B respondia os mesmos utilizando a classificação. O problema com essa solução é que um usuário A e um usuário B podem fazer diferentes consultas para o mesmo tópico, obtendo assim resultados diferentes. O que poderia ser resolvido caso determinássemos quais deveriam ser as consultas utilizadas. Porém, isso restringiria ainda mais o conjunto de testes à nossa visão de como funciona e o que é útil em um sistema de busca; por isso, essa solução não foi adotada.

[58] Como dito no Capítulo 6 o protótipo Leva-e-traz busca apenas 20 resultados devido ao longo tempo gasto com download e transformação para texto dos mesmos para que possam ser classificados. Utilizá-lo em uma avaliação com usuários com um número maior de resultados, só seria viável com uso de cache. Porém, para isso teríamos que determinar não apenas os tópicos, mas também as consultas, o que consideramos ser um prejuízo maior do que a limitação do número de resultados.

Quadro 21 – Tópicos de busca utilizados na avaliação

(1) O MORRO DOS VENTOS UIVANTES. Encontre páginas que vendam o livro ou o filme O MORRO DOS VENTOS UIVANTES. Páginas que descrevem o livro ou o filme, mas não oferecem os mesmos para venda, não atendem a consulta. **(2) BODE E CARNEIRO PARA A PARAPSICOLOGIA**. Encontre páginas que definam bode e carneiro segundo a parapsicologia. Páginas que falem de bode e carneiro na parapsicologia, mas não digam o significado destes dois termos, não atendem a consulta. **(3) CAUSAS DE INCÊNDIOS DOMÉSTICOS**. Seu objetivo é encontrar quais as principais causas de incêndios no lar. Uma página atende a consulta se mencionar, pelo menos, uma causa de incêndio em residências privada. **(4) PREVENÇÃO DA RAIVA EM SERES HUMANOS**. O objetivo é encontrar páginas que discutam métodos de prevenção da raiva em pessoas. Uma página para atender a consulta, deve citar pelo menos uma maneira para a prevenção da forma humana da raiva. **(5) MISÉRIA NA ÁFRICA**. O objetivo é encontrar páginas que tragam informações sobre a miséria na África. Páginas atendem a consulta caso mencionem quaisquer aspectos, tais como causas ou estatísticas sobre a Miséria na África. Notícias sobre eventos para discutir a Miséria na África, que não tragam informações sobre a mesma, não atendem a consulta. **(6) MENSALÃO**. O objetivo é encontrar páginas que definam o que é e como fazer o recolhimento complementar (mensalão). Notícias ou comentários sobre o escândalo do mensalão[59] não atendem a esta consulta.

Os dois primeiros tópicos fazem parte da lista de 42 pares de consultas/objetivos de usuários levantadas por Aires & Aluísio (2003); os tópicos 3 e 4 fazem parte da lista de 50 tópicos da coleção CHAVE[60] para o CLEF 2004; e os dois últimos tópicos foram criados especificamente para essa avaliação. A seleção dos tópicos 1 a 4 deu-se através da submissão dos tópicos de ambas coleções ao Google e ao Alltheweb para que selecionássemos tópicos para os quais as respostas devolvidas atendessem a mais de um tipo de necessidade. A maioria dos tópicos foi eliminada porque, quando pesquisados, eram apresentados apenas notícias ou apenas panoramas. Para o primeiro tópico, são retornados críticas, lista dos filmes produzidos pelo diretor do filme, produção bibliográfica da autora do livro, sinopses, resumo do livro e páginas que vendem o filme ou o livro. Para o segundo, são retornados o artigo de Aires & Aluísio (2003); páginas de institutos de parapsicologia; bibliografias, notícias, artigos, e blogs sobre outros assuntos que contém os termos "bode", "carneiro" e parapsicologia", mas que são sobre outros assuntos; páginas de livrarias, de livros com anúncios de livros sobre outros assuntos e um artigo sobre a pesquisa parapsicológica na metapsíquica, que realmente define os termos.[61] Para o terceiro tópico, foram encontradas, além de páginas que realmente mencionam causas de incêndios domésticos, páginas que utilizam os termos da consulta, mas falam de

[59] Em 2005 descobriu-se que parlamentares recebiam propina mensal para apoiar o governo em votações. Tal escândalo ficou conhecido como "Mensalão".
[60] http://www.linguateca.pt/CHAVE/
[61] Termos cunhados por Gertrude Schmeidler para designar pessoas propensas ou refratárias à experiência psi, denominando as primeiras de carneiros e as últimas de bodes.

outros temas como combustão do corpo humano e causas de danos em computadores, além de notícias. Para o quarto tópico, foram encontradas, além de páginas relevantes, panoramas, definições e notícias. Por exemplo, definição do que é uma vacina anti-rábica e uma notícia sobre o preparo de uma equipe para atuar na prevenção da raiva. Já os últimos dois casos foram criados a partir de temas para os quais muitos documentos irrelevantes seriam retornados. A pesquisa por "Miséria + África" ou "Miséria na África", por exemplo, tinha apenas um resultado relevante, todos os outros mencionavam as duas palavras, mas não discutiam o tema; eram todos notícias sobre o "Live 8"[62] ou blogs apoiando essa iniciativa. O mesmo ocorreu para a pesquisa "mensalão", que, na semana em que foram realizados os experimentos, tinha apenas um documento relevante, pois todos os outros falavam sobre o escândalo da compra de voto de parlamentares apelidado por mensalão e não do tipo de recolhimento complementar que tem o nome de mensalão.

Para cada tópico, os usuários deveriam informar: as consultas utilizadas, a relevância dos resultados e se a classificação dos resultados os induziu a erros em seus julgamentos sobre a relevância dos mesmos. Essa última pergunta foi feita para cada tópico para nos poupar do esforço de verificar a classificação para cada um dos 1200 resultados[63], ainda que fosse interessante verificar a taxa de acerto dos classificadores em novos dados. Essa simplificação não nos causa problemas, uma vez que o objetivo dessa avaliação é verificar se, com as taxas de acerto estimadas para os classificadores no Capítulo 8, a classificação auxilia o usuário em seus julgamentos.

Os julgamentos de relevância são o meio proposto nesta avaliação para analisarmos o esforço de busca de um usuário, comparando-se quantos resultados irrelevantes são vistos durante sua procura em um sistema tradicional e em um sistema com a classificação em necessidades de busca. O que é feito de três formas, verificando-se para os resultados classificados ou não: se o primeiro resultado é relevante, se o primeiro resultado que o usuário pensa ser relevante é relevante e comparando-se o número de resultados que o usuário pensa ser relevantes e realmente são, dados apenas o título e resumo, ou dada também a classificação.

[62] Série de concertos que aconteceram simultaneamente em diferentes cidades do mundo para mostrar a miséria na África e pressionar os líderes do G8, que se reuniram entre 6 e 8 de julho de 2005 na Escócia, a ajudarem a atacar esse problema.
[63] Número máximo de resultados considerando-se 6 tópicos, 20 resultados por tópico e os 10 usuários.

10.2 Resultados

A apresentação das sete necessidades com seus exemplos, na página de *Help* do Leva-e-traz, teve resultados positivos para a clareza da taxonomia de sete necessidades, uma vez que o tipo de necessidade foi escolhido de forma errada em apenas 1 das 30 consultas realizadas utilizando-se a *tab* necessidades. Um usuário selecionou o tipo 2 "explique como fazer" para o tópico 2 "definição de bode e carneiro para a parapsicologia".

Em 29 das 30 consultas utilizando necessidades, o primeiro resultado era um resultado relevante e, nas 30 consultas não classificadas, o primeiro resultado era relevante para 18 das consultas.

Solicitou-se ainda que, dadas apenas as informações apresentadas na tela, o usuário clicasse em um resultado que aparentemente atendesse melhor a consulta. Para a classificação em necessidade, em 20 das 30 consultas, o resultado escolhido foi realmente relevante e, para os resultados sem classificação, em 13 das 30 consultas.

Havia sido solicitado também que, após os usuários marcarem todos os resultados que achavam serem relevantes, os resultados fossem visitados para que se verificasse se eram mesmo relevantes. Porém, esses números não são mostrados aqui, pois encontramos um problema grave durante a análise dos dados: quatro dos dez usuários disseram ter encontrado um número maior de documentos relevantes do que o que realmente existia para suas consultas. Por exemplo, um dos usuários disse ter selecionado 13 dos 19 documentos inicialmente retornados para a consulta "mensalão" como relevantes e que efetivamente 6 eram relevantes. Porém, no mesmo dia, a mesma consulta tinha tido todos os seus resultados verificados e existia apenas 1 resultado relevante. O mesmo usuário disse ter encontrado 8 documentos relevantes para uma consulta sobre "miséria na África" que também havia sido feita anteriormente e para a qual havia apenas 1 documento relevante. Os motivos podem ser: (a) que a análise da relevância não tenha sido feita com cautela ao visitar os resultados, por exemplo, ou (b) por demasiada autoconfiança sobre o que seria relevante (o usuário acima referido informou não utilizar a classificação em nenhum de seus julgamentos), ou (c) por displicência com o experimento devido a demora no tempo de resposta, que é de cerca de três minutos para algumas das consultas (o

usuário, acima referido em particular, colocou como comentário em seu questionário que a consulta sobre mensalão demorou muito). As consultas digitadas para cada tópico pelos usuários são mostradas no Quadro 22.

Quadro 22 – Consultas digitadas pelos usuários para cada um dos seis tópicos

(4) "morro dos ventos uivantes" (2) O morro dos ventos uivantes (1) Compra "o morro dos ventos uivantes" (1) "o morro dos ventos uivantes" preço (1) filme +livro + "o morro dos ventos uivantes" (1) "morro dos ventos uivantes" R$
(4) Bode carneiro parapsicologia (2) Bode e carneiro para a parapsicologia (1) Parapsicologia bode carneiro (1) Parapsicologia +bode e carneiro (1) Bode +carneiro +parapsicologia (1) Bode carneiro parapsicologia psicologia
(7) Causas de incêndios domésticos (1) Incêndios domésticos causas (1) "incêndios domésticos" causas (1) "causas de incêndios" +domésticos +lar
(5) Prevenção da raiva em seres humanos (1) Métodos prevenção raiva (1) Prevenção da raiva +seres humanos (1) Raiva "seres humanos" prevenção (1) "raiva canina" prevenção (1) prevenção prevenir raiva seres humanos
(5) Miséria na África (1) Miséria (1) Miséria + África (1) África miséria (1) causas + "miséria na África" (1) causas estatísticas miséria fome África
(7) Mensalão (1) "recolhimento complementar" (1) mensalão + "como funciona" (1) mensalão repasse conta recolhimento complementar

Perguntou-se, também, se a classificação havia induzido a julgamentos de relevância equivocados, o que aconteceu para seis das trinta consultas (20%)[64], e se o usuário havia clicado em resultados relevantes que não clicaria não fosse a classificação, o que aconteceu em dez das trinta consultas (33%).

As últimas perguntas do questionário foram sobre a utilidade da classificação em necessidades e sobre sugestões de inclusão de outros tipos de necessidades. Três

[64] Um dos usuários foi induzido a julgamentos errados para dois tópicos e outros quatro usuários foram induzidos a julgamentos errados para um tópico. Porém, não sabemos para quantos dos resultados apresentados em cada um desses tópicos a classificação estava incorreta.

dos dez usuários disseram que o esquema não era útil, os mesmos três reportaram para todas as consultas que não utilizaram a classificação para efetuar seus julgamentos de relevância. Dois desses fazem parte dos quatro mencionados anteriormente que encontraram um número maior de documentos relevantes do que realmente existia.

Apenas duas pessoas teceram comentários sobre o esquema de classificação em necessidades. O primeiro sugerindo que fossem dados mais exemplos para cada tipo de necessidade. E o segundo sugerindo dois tipos novos, encontrar: (i) "informações críticas e subjetivas, opiniões próprias, editoriais, resenhas, prefácios de livros (como blogs e páginas pessoais)"; (ii) "informação estatística (índices como o IDH)".

10.3 Considerações sobre os resultados

No planejamento desse experimento, considerou-se o fato de que, por um usuário estar habituado com um dado sistema, pudesse haver resistência quanto ao uso da informação de classificação em necessidades. Por isso, para cada tópico em que se deveria usar a classificação, perguntou-se ao usuário se o mesmo havia ou não utilizado essa informação em seu julgamento. Porém, não se contou com a possibilidade de os julgamentos de relevância serem feitos superficialmente ou por hábito ou por displicência com o experimento, uma vez que participaram do experimento apenas voluntários[65]. O que levanta a questão de como medir o esforço do usuário de forma precisa. Quando essa parte do experimento foi elaborada, pensou-se que seria difícil aplicá-la a um grupo grande de usuários, com pessoas com diversas formações e idades para que fosse possível generalizar os resultados para usuários em geral da Web. Porém, essa medida mostrou-se não ser eficiente nem mesmo em um grupo seleto e pequeno.

Em experimentos futuros, acreditamos que a avaliação possa ser mais precisa caso: (a) avaliemos cada um dos resultados verificando a taxa de acerto assim em novos dados; (b) verifiquemos, nos casos em que o usuário diz que a classificação o induziu a erros, quais foram os erros cometidos; (c) avaliemos o mesmo conjunto de

[65] Oito deles trabalham com pesquisas científicas de algum tipo e dois trabalham com testes sistemáticos de sistemas, ou seja, todos têm de certa forma conhecimento sobre a importância do rigor científico para a pesquisa.

resultados com outros usuários; (d) questionemos os usuários sobre o interesse a priori nos tópicos utilizados no teste; (e) proponhamos ao usuário que faça uso da classsificação para 1 ou dois tópicos livres pelos quais se interesse; e (f) analisemos estatisticamente se a melhoria com a classificação é devida à qualidade do método ou a termos avaliados os dados com diferentes usuários.

Para avaliar os mesmos tópicos com outros usuários, pode-se, dados os resultados para as consultas digitadas pelos usuários reunidas em um conjunto de teste, solicitar aos novos usuários que os julguem. Assim, pode-se verificar se outros usuários fariam as mesmas escolhas quanto à relevância/irrelevância de um resultado.

Questionando-se os usuários se fariam perguntas como os tópicos de teste, pode-se determinar seu interesse a priori nos tópicos que fazem parte do teste. Assim, poderíamos chegar a hipóteses mais embasadas sobre o motivo de usuários terem julgado documentos irrelevantes como relevantes.

É importante, ainda, que, em futuras avaliações, sejam feitas análises mais precisas sobre a qualidade da classificação, para verificarmos se a melhoria do esforço é devido ao método ou aos diferentes grupos de usuários, ou seja, verificar se as pessoas que fizeram os primeiros três tópicos não tinham um perfil diferente das que fizeram os outros três com ajuda da classificação. Para isso, um modelo a seguir poderia ser o trabalho de He *et al*. (2004), embora noutra área, a da resposta a perguntas. Os autores, com 16 perguntas e 8 usuários diferentes, além de medirem/analisarem diferentes graus de dificuldade e a correlação entre o tempo/número de procuras e a qualidade da resposta, capturaram em vídeo as ações dos usuários para depois fazerem uma análise mais fina do que de fato acontecia. Contudo, não foram capazes de obter resultados estatisticamente relevantes para a sua comparação, o que significa, muito provavelmente, que experimentos como os descritos nesta tese necessitam de muito mais usuários e tópicos para poderem produzir esse tipo de resultados.

11. Conclusão

"Education in 2020 will be about learning how to acquire and use information, because such skills will be far more important to future generations than the learning of 'facts'." IEE (2000)[66]

Como visto no Capítulo 1, o tópico/assunto de documentos é a característica principal utilizada por sistemas de recuperação nas últimas décadas. Entretanto, a forma como essa informação é apresentada é uma questão negligenciada por grande parte dos sistemas de informação (Rauber & Müller-Kogler, 2001). Uma maneira teoricamente simples de dizer aos usuários como uma informação é apresentada por um documento é a classificação automática de textos segundo gêneros que, como visto no Capítulo 5, tem uma boa taxa de acerto. Porém, no caso da Web, apenas classificar os resultados segundo gêneros pode não ser suficiente para esclarecer o usuário sobre a relevância, ou não, de um documento para sua consulta, por diversos motivos:

- Gênero é um conceito de certa forma subjetivo, tanto porque a lista de categorias pode ser determinada de forma diferente (veja os diferentes exemplos de taxonomias apresentados no Capítulo 5), quanto porque o que pode ser classificado como pertencente a uma dada categoria varia de acordo com a nossa interpretação do que é a categoria. Se pesquisadores especializados em estilometria têm diferentes visões sobre o que são gêneros, quantas não são seriam as interpretações de usuários da Web;
- Projetistas não seguem regras quanto ao conteúdo para desenvolver páginas da Web, documentos podem cobrir diferentes gêneros. O que pode acontecer porque um mesmo documento pode ser classificado segundo dois ou mais tipos de gêneros, ou por conter seções de diferentes gêneros;
- São vários os novos gêneros da Web, muitos com características muito semelhantes a outros gêneros, veja o caso de *FAQs*, mensagens de discussões e entrevistas, todos formados por perguntas e respostas;
- Finalmente, as consultas de um usuário podem ser atendidas por mais de um gênero. Por exemplo, se deseja encontrar uma determinada definição, essa pode

[66] http://www.iee.org/News/PressRel/z26may2000.cfm

ser dada por trabalhos científicos, jornalísticos ou, ainda, em página de curiosidades.

Conclui-se, então, que a classificação em gênero é uma informação complementar ao assunto bastante interessante para auxiliar o usuário a julgar a relevância ou não de um documento. Porém, ainda mais interessante seria se informássemos se uma página atende a determinados tipos de necessidades, seja isso feito através de um gênero ou mais. Porém, os porquês, objetivos de busca de um usuário, são uma questão ainda mais negligenciada pelos sistemas de busca na Web atuais. Em nossa revisão da literatura, encontramos trabalhos que discutem taxonomias de necessidade de busca, como o de Broder (2002) e o de Rose & Levinson (2004) e trabalhos que classificam consultas de usuários segundo necessidades que esses querem ver atendidas, por exemplo, o de Kang & Kim (2003). Porém, o trabalho apresentado nesta tese é o primeiro a classificar documentos como atendendo ou não a necessidades de busca. Tarefa que se provou poder ser bem realizada para textos da Web em português. Obteve-se 79,38% de taxa de acerto para a taxonomia de sete necessidades e 91,19% para a distinção entre serviços e informações, duas das categorias da taxonomia de consultas de Broder (transacional e informacional), o que é suficiente para atendê-la, uma vez que consultas do tipo navegacional podem ser atendidas utilizando-se técnicas como a análise de *links*, utilizada na tarefa de busca de home pages (*home page finding*) do TREC. Desenvolveu-se, também, com sucesso, a classificação segundo necessidades personalizadas, com uma taxa de acerto entre 65,5% e 89,91%.

11.1 Contribuições

São cinco as contribuições principais deste trabalho:

- Investigou-se com sucesso a classificação de documentos da Web segundo necessidades de busca; para isso diferentes conjuntos de marcadores de estilo foram também investigados;
- Mostrou-se indicações de que usuários estariam interessados em formas de classificação personalizadas, mesmo que tenham que contribuir para a construção das mesmas;

- Investigou-se o uso de algoritmos de aprendizado de máquina e marcadores de estilo para uma nova aplicação de estilometria, a classificação segundo necessidades;
- Mostrou-se um caminho a seguir em avaliações em que o sistema é analisado sob o ponto de vista do usuário;
- Apesar de ser um trabalho inovador para qualquer língua, trabalhou-se com o português, pois acreditamos que mais importante do que obter destaque internacional trabalhando-se com o inglês é primeiro contribuir para a não exclusão, na Web, de falantes de português.

Como contribuições indiretas:

- Disponibilizou-se na Web os *corpora* classificados segundo necessidades e necessidades personalizadas, que poderão ser utilizados por outros pesquisadores tanto para trabalhos de classificação quanto para trabalhos de PLN e de lingüística computacional, para, por exemplo, testar sistemas que usam textos da Web ou fazer análises lingüísticas;
- Disponibilizou-se, também, o protótipo Leva-e-traz, apesar do mesmo não ter sido concebido para ser um produto. Seu código está disponível e inclui um programa para transformar documentos html, doc e pdf para txt e scripts para cálculo das *features* utilizadas;
- Relatou-se, nessa tese e em artigos publicados, a lista completa de marcadores utilizados. O fato de termos deixado disponíveis os recursos utilizados torna possível a replicação dos estudos; as listas de marcadores, em particular, possibilitam a replicação dos experimentos mesmo por pesquisadores que não conhecem a língua portuguesa.
- Demonstrou-se que marcadores de estilo similares aos utilizados para outras línguas também distinguem com sucesso gêneros em português. Obtivemos uma taxa de acerto de 94,87% com o *corpus* com 5 gêneros, sendo os textos dos gêneros informativos e instrucionais extraídos da Web. A classificação em gêneros não só é uma informação complementar útil para a busca na Web, como poderia ser importante para várias aplicações de PLN, como, por exemplo, para aumentar a taxa de acerto de etiquetadores morfossintáticos, etiquetadores

sintáticos e desambigüizadores de sentido (muitos sentidos são restritos a textos de determinado estilo, como, por exemplo, formal e informal).

11.2 Limitações

Como toda nova idéia, nossa abordagem tem algumas limitações que precisam ser investigadas: tempo de resposta, universalidade do conceito das sete necessidades, como convencer usuários sobre a credibilidade da classificação, e qual a taxa de acerto mínima para que a classificação seja útil.

No caso do Leva-e-traz, a classificação dos documentos é feita no momento da consulta, devido a tempo de *download* e transformação para texto dos documentos; o tempo de resposta chega perto de 4 minutos. Acreditamos que essa limitação é apenas de nosso protótipo, uma vez que em máquinas de busca a classificação poderia ser feita previamente no momento da indexação. Ainda assim, seria interessante investigar, por exemplo, se a classificação dos textos poderia ser feita utilizando-se apenas partes deles, ou se é imprescindível que se utilize todo o texto de um documento.

Como dito anteriormente, gerar um esquema de classificação é de certa forma subjetivo, porque podem existir diferentes pontos de vista sobre o que pode ser coberto por uma categoria. Apesar de não propormos um esquema de gêneros, como foi mostrado no primeiro questionário, existem outras interpretações possíveis sobre o que uma classe pode conter. Veja o caso do usuário que considerou que poderia encontrar uma lista de hotéis utilizando a necessidade serviços (Capítulo 7). Aparentemente, pelos resultados mostrados no Capítulo 10, o fato de termos apresentado exemplos, junto à lista de necessidades no arquivo de *help*, solucionou esse problema. Porém, nossa população de estudo, 10 usuários, é muito pequena para ser considerada uma amostra razoável de usuários da Web, pois na Web estaremos lidando com pessoas de diferentes formações, culturas e que podem interpretar as mesmas palavras de maneiras diferentes. Fica sem resposta a questão de se seria possível ter uma lista de exemplos tão exaustiva que não deixasse dúvidas para os diversos tipos de usuários. Essa questão é muito importante, pois, ainda que interpretemos a necessidade do usuário através de sua consulta automaticamente, ao

invés de solicitarmos que ele escolha um tipo de classificação, o usuário teria de entender os nomes dados a cada tipo de necessidade para fazer uso dela em seu julgamento. No caso dos exemplos, seria importante estudar uma forma de mostrá-los em momentos diferentes da busca, pois sabemos que usuários podem não ler informações de ajuda.

No Capítulo 10, reportamos o caso de quatro usuários que "encontraram" para suas consultas um número maior de documentos relevantes do que o que realmente existia; três desses usuários informaram para todas as consultas que não utilizaram a informação de classificação em necessidades apresentada na tela. O que nos fez questionar se isso era devido à maneira com que o usuário está acostumado a fazer suas buscas ou se, por exemplo, estava relacionado à qualidade das outras informações apresentadas, como, por exemplo, o resumo do documento. Isto é, se, dado um resumo pobre, o usuário confiaria na classificação, que ela estava correta apesar de sua intuição ser diferente. Essa questão também precisa ser melhor avaliada com novos estudos com usuários. Caso se confirme, a classificação só seria útil se todo o sistema de busca se tornasse melhor. Outra questão, não avaliada nesse trabalho e que também deve ser analisada em futuros experimentos, é qual a taxa de acerto mínima para que um usuário confie e considere o sistema útil.

11.3 Trabalhos futuros

Este projeto poderia ser estendido de várias maneiras, dentre elas: averiguando-se a relação entre tamanho de texto e taxa de acerto de classificação; explorando-se mais marcadores estilísticos e algoritmos para geração de classificadores; construindo-se um *corpus* que pudesse ser utilizado por variados trabalhos de estilometria; investigando-se a utilidade de marcadores semelhantes aos utilizados neste projeto para a classificação de textos em inglês ou outras línguas; e estudando métodos de classificação incrementais para a classificação em necessidades personalizadas.

11.3.1 Relação entre tamanho do texto e taxa de acerto

Biber (1993) diz que um tamanho de texto de 1000 palavras é adequado para representar as distribuições de várias características lingüísticas com funções

estilísticas. Stamatatos *et al* (2000b) dizem que, aparentemente, em seus experimentos, os resultados foram mais confiáveis para textos com mais de 1500 palavras. Nos *corpora* de necessidades utilizados nesse trabalho, devido à grande variação das fontes, havia textos de tamanhos variados, alguns com menos de 100 palavras. Um estudo futuro importante é verificarmos a importância do tamanho dos textos para nossos esquemas de classificação, através de experimentos com *corpora* compostos cada um por textos de tamanhos semelhantes, 100, 500, 1000, 1500, 2000 ou mais palavras.

11.3.2 Marcadores estilísticos e algoritmos para classificação

Em nossos últimos experimentos realizados com marcadores estruturais e sintáticos, não foi identificada melhoria na taxa de acerto, o que intuitivamente é estranho. É importante verificar em experimentos futuros se isso não aconteceu por *overfitting*. Se for o caso, o ideal seria trabalhar com um *corpus* maior e testar esses conjuntos de marcadores com algoritmos mais resistentes a *overfitting* como redes neurais. Outra estratégia que poderia ser utilizada caso o problema seja *overffiting*, seria gerar um classificador por conjunto de marcadores e combinar suas saídas de classificação utilizando, por exemplo, votação pela maioria (*majority voting*). Outra possibilidade é, ainda, investigar uso de técnicas para seleção de *features*, como *information gain*, uso de z-scores etc., para selecionarmos as *features* mais discriminantes, dados todos os conjuntos de marcadores.

Outras extensões para o trabalho de classificação em necessidades seriam o uso de outros marcadores, como funções para calcular a legibilidade de textos (*readability*)[67], outros marcadores estruturais, como tabela e mudanças em fontes em documentos que não são html; entidades mencionadas; palavras funcionais extraídas a partir de *corpora* representativos do português. Além dessas, investigar possíveis relações retóricas (Marcu, 2000), identificadores semânticos e pragmáticos que pudessem ser utilizados como marcadores, mas que não fossem dependentes de domínio.

[67] http://www.readability.info/info.shtml

11.3.3 Corpus padrão para testes

Argamon & Levitan (2005) sugerem que a comunidade de estudos sobre atribuição de autoria deveria trabalhar em conjunto na criação de um grande conjunto de *corpora* e tarefas de avaliação, para permitir comparações mais rigorosas e padronizadas das abordagens para determinação de autoria. Advogam também a favor de conjuntos de teste maiores, para avaliar a utilidade de diferentes conjuntos de *features* e técnicas. Em seu trabalho, os resultados utilizando palavras freqüentes foram melhores do que o resultado utilizando colocações, o que contradiz os resultados de outros trabalhos relacionados. Afirmam, baseados nessa informação que mudar a escala do conjunto de dados pode afetar os resultados de diferentes técnicas, levando a conclusões diferentes de acordo com a escala.

No nosso caso, seria também muito interessante que tivéssemos *corpora* com muitos textos. O esforço em conjunto não teria de ser necessariamente utilizando as seis necessidades levantadas neste trabalho. Poderia ser, por exemplo, em funções mais gerais, tais como estilo público, estilo científico, estilo jornalístico, estilo de comunicação do dia-a-dia e estilo literário (Michos *et al*, 1996). Os textos, sendo classificados segundo tais funções, poderiam ser adaptados para diferentes tarefas como a nossa. Utilizar um número de textos de treinamento bem maior que o número de *features*, reduziria as chances de *overfitting* do modelo para os dados de treinamentos, aumentando a credibilidade dos resultados. Mas, seria interessante que esses *corpora* fossem criados com textos da Web, já que suas características podem ser bem diferentes dos textos em outras mídias, como, por exemplo, os textos jornalísticos. *Corpora* como esses poderiam ser utilizados não só em trabalhos de classificação de textos, mas também de lingüística e literatura. Os *corpora* poderiam ser utilizados em vários estudos de estilometria, como no uso dos marcadores levantados para clusterização dos textos, classificação segundo outros critérios, como se um texto emite opinião, se é uma narrativa e textos com uma visão positiva ou negativa dos assuntos.

11.3.4 Uso de marcadores estilísticos para a classificação em necessidades de textos em outras línguas

Em Aires *et al* (2005a) apresenta-se um experimento para a classificação de textos de direito em inglês em textos para leigos ou para especialistas. Estudos semelhantes do uso de marcadores estilísticos poderiam ser feitos tanto para a classificação em necessidades personalizadas, como nesse caso, quanto para a classificação em sete necessidades, para outros idiomas, já que, teoricamente, os mesmos seriam universais. No caso do inglês, poder-se-ía, por exemplo, utilizar diretórios como Yahoo para simplificar a seleção de páginas para o *corpus* de treinamento.

11.3.5 Treinamento incremental

Seria interessante estudar, para a classificação em necessidades personalizadas, algoritmos para os quais o treinamento do classificador pudesse ser incremental. Assim, o usuário poderia fornecer, aos poucos, os dados de treinamento, por exemplo, à medida que navegasse na Web e visse uma nova página que serviria como exemplo para seu *corpus* de treinamento.

11.4 Considerações finais

O estilo é questão presente em qualquer texto, seja estilo pessoal do autor, estilo imposto pelo meio, pelo gênero ou pela audiência, ou uma combinação desses. Neste projeto, fez-se uso de marcadores estilísticos relacionados a tipos de texto para que se pudesse classificar resultados de busca segundo intenções de usuários, o que, sem sombra de dúvidas, apesar das limitações, demonstrou ser um caminho promissor a ser investigado. Como pôde ser visto pelos exemplos de extensões para este projeto dados nas seções anteriores, e esses são mesmo apenas alguns exemplos, são muitos os caminhos que podem ser seguidos na investigação de estilo para melhoria da qualidade dos resultados de uma busca na Web. Note-se que a lista de trabalhos futuros foi feita considerando-se apenas características de estilo referentes a tipos de texto e como aplicação a busca na Web, se tivéssemos considerado todo o PLN ou RI, muito mais haveria a dizer.

Estilo é característica importantíssima de textos, tanto quanto conteúdo, pois de diversas maneiras fornece indicações sobre o que pode ser um indicador de qualidade para um dado usuário em um dado momento. Por isso, é imprescindível que as comunidades de processamento de linguagem natural e de lingüística para português apostem mais fortemente em pesquisas sobre usos de estilo e características estilísticas de textos na nossa língua. Esta tese é original por considerar, pela primeira vez, na classificação de resultados de busca, os motivos de um usuário efetuar uma dada pesquisa e isso é feito através do uso de marcadores estilísticos. Porém, poucos foram os estudos que vimos em nossa revisão bibliográfica para português de estilometria, considerando-se usos já estudados para outras línguas, tanto em aplicações em geral quanto em estudos teóricos.

Quanto ao futuro, estamos plenamente convencidos de que, em menos de dez anos, todos os motores de busca irão recorrer a personalização dos resultados para facilitar a exploração destes pelo usuário. O que pode ser feito com a ajuda de marcadores estilísticos. Daqui a dez anos, nenhum usuário imaginará o esforço de separar o trigo do joio a que somos de momento obrigados, graças aos esforços recentes de personalizar os sistemas de busca, manipulação e gerência de informação, dentre eles o tratamento de estilo dos resultados da procura. Este é apenas o primeiro passo nessa direção.

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[68]

(Aiken & West, 1996) Aiken, Leona S.; West, Stephen G. Multiple Regression. Sage Publications Inc, 1996, 224 p.

(Aires & Aluísio, 2002) Aires, R.; Aluísio, S. (2002). Eu falo português. E daí? Poster in IHC 2002 – 5th Symposium on Human Factors in Computer Systems, outubro de 2002, Fortaleza - Brasil.

(Aires & Aluísio, 2003) Aires, R.; Aluísio, S. (2003). Como incrementar a qualidade dos resultados das máquinas de busca: da análise de *logs* à interação em português. Revista Ciência da Informação, vol 32, n. 1, p. 5-16, janeiro/abril de 2003.

(Aires & Santos, 2002) Aires, R.; Santos, D. (2002). Measuring the Web in Portuguese. In EuroWeb 2002 conference, p. 198-199, dezembro de 2002, Oxford - Inglaterra.

(Aires *et al*, 2003) Aires, R.; Aluísio, S.; Quaresma, P.; Santos, D.; Silva, M. (2003). An initial proposal for cooperative evaluation on information retrieval in Portuguese. In PROPOR 2003 – 6th Workshop on Computational Processing of the Portuguese Language, p. 227-234, junho de 2003, Faro - Portugal. (c) Springer-Verlag.

(Aires *et al*, 2004a) Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) Which classification algorithm works best with stylistic *features* of Portuguese in order to classify Web texts according to users' needs? Relatório técnico nº 241, outubro de 2004, ICMC/USP.

(Aires *et al*, 2004b) Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) What is my Style? Using Stylistic *features* of Portuguese Web Texts to classify Web pages according to Users'Needs. In LREC 2004, p. 1943-1946, maio de 2004, Lisboa - Portugal.

(Aires *et al*, 2005a) Aires, R.; Aluísio, A.; Santos, D. (2005) User-aware page classification in a search engine. A ser publicado nos Proceedings of 2005 SIGIR Workshop on Textual Stylistics in Information Access. SIGIR, agosto de 2005, Salvador – Brasil, 8 p.

(Aires *et al*, 2005b) Aires, R.; Santos, D.; Aluísio, A. (2005) "Yes, user!": compiling a *corpus* according to what the user wants. *Corpus* Linguistics 2005, julho de 2005, Birmingham – Inglaterra, 14 p. Disponível em www.corpus.bham.ac.uk/PCLC .

(Aires *et al*, 2005c) Aires, R.; Aluísio, S. (2005) "As avaliações atuais de sistemas de busca na Web e a importância do usuário". A ser publicado em Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.

[68] Todos os links estavam ativos em 01/08/2005.

(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.

(Allen, 1990) Allen, R. User models: Theory, method and practice. International Journal of Man-Machine Studies, 32, 1990, p. 511-543.

(Aluísio *et al*, 2003) Aluisio, S.; Pinheiro, G.; Finger, M.; Nunes, M.G.V.; Tagnin, S.E. The Lacio-Web Project: overview and issues in Brazilian Portuguese *corpora* creation. In Proceedings of *corpus* Linguistics (2003), Lancaster, Inglaterra. v. 16, p. 14-21.

(Arampatzis *et al*, 1999) Arampatzis, A. T.; van der Weide, Th. P.; van Bommel, P.; Koster, C. H. A. 1999. Linguistically Motivated Information Retrieval. University of Nijmegen, Relatório técnico CSI-R9918. 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.

(Argamon & Levitan, 2005) Argamon, Shlomo; Levitan, Shlomo. Measuring the Usefulness of Function Words for Authorship Attribution. Proceedings of ACH/ALLC Conference 2005 in Victoria, BC, Canadá, junho de 2005. Disponível em http://lingcog.iit.edu/doc/paper_162_argamon.pdf

(Argamon *et al*, 1998) Argamon, Shlomo; Koppel, Moshe; Avneri, Galit. Routing documents according to style. In First International Workshop on Innovative Information Systems, 1998. Disponível em: http://www.idt.ntnu.no/~monica/iiis-98/papers/argamon.ps

(Argamon *et al*, 2003) Argamon, Shlomo; Koppel,Moshe; Fine, Jonathan; Shimoni, Anat Rachel. Gender, Genre, and Writing Style in Formal Written Texts. Text (Berlin, Germany), 23, 3, 2003, p. 321-346. Disponível em http://www.cs.biu.ac.il/~koppel/papers/male-female-text-final.pdf

(Baeza-Yates & Ribeiro-Neto, 1999) Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, 1999, 544 p.

(Baljko & Hirst, 1999) Baljko, Melanie; Hirst, Graeme. The importance of subjectivity in computational stylistic assessment. Text Technology, 9(1):5-17. 1999. Disponível em http://www.cs.utoronto.ca/~melanie/pubs/TT-99.ps

(Bar-Ilan & Peritz, 2005) Bar-Ilan, Judit; Peritz, Bluma C. Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of informetrics. Journal of the American Society for Information Science and Technology. Volume 55, Issue 11, p. 980-990, setembro de 2004. Disponível em http://www3.interscience.wiley.com/cgi-bin/fulltext/107640586/HTMLSTART.

(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.

(Belkin & Croft, 1992) Belkin, N. J.; Croft, W. B. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 1992. Disponível em www.ischool.utexas.edu/~i385d/ readings/Belkin_Information_92.pdf

(Bharat & Broder, 1998) Bharat, K.; Broder, A. A technique for measuring the relative size and overlap of public Web search engines. Proceedings of the 7th International World Wide Web Conference: Vol. 30. Computer Networks and ISDN Systems, p. 379-388, 1998. Retrieved December 15, 2003.

(Bhatt *et al*, 2004) Bhatt, Khelan; Evens, Martha; Argamon, Shlomo. Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions. Proceedings of the 26th Annual Meeting of the Cognitive Science Society, agosto de 2004. Disponível em: http://lingcog.iit.edu/doc/bhattevensargamonsubmit.pdf

(Biber & Finegan, 1994) Biber, D.; Finegan, E. Sociolinguistic perspectives on register. Oxford, Oxford University Press, 1994.

(Biber, 1986) Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. Language, 62(2), p. 384-413.

(Biber, 1988) Biber. Douglas. 1988. Variation across Speech and Writing. Cambridge University Press. Cambridge. England.

(Biber, 1992) Biber. Douglas. 1992. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding. Computers in the Humanities, 26(5-6), p. 331-347.

(Biber, 1993) Biber, D. Using Register-Diversified *corpora* for General Language Studies. Computational linguistics, 19/2, p. 219-241, 1993.

(Biber, 1995) Biber. Douglas. 1995. Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press. Cambridge. England.

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

(Bilal & Kirby, 2001) Bilal, Dania; Kirby, Joe. Differences and similarities in information seeking: children and adults as Web users. Information Processing and Management: an International Journal, 2001, 38 (5), p. 649-670.

(Bowles, 1998) Bowles, Mark D. The Information Wars: Two Cultures and the Conflict in Information Retrieval, p. 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

(Bretan *et al*, 1998) Bretan, Ivan; Dewe,Johan; Hallberg, Anders; Wolkert,Niklas; Karlgren, Jussi. 1998. Web-Specific Genre Visualization". In Proceedings of the Webnet World Conference on the WWW and Internet, Orlando, Florida.

(Brill, 1994) Eric brill. Some advances in transformation-based parts of speech tagging. In AAAI, 1994.

(Broder, 2002) Broder, A. "A Taxonomy of Web Search", SIGIR Forum 36 (2), Fall 2002, p.3-10.

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

(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, fevereiro de 2002. Disponível em: http://www.dgzero.org/fev02/Art_05.htm

(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.

(Calado, 1999) Calado, P. (1999) The WBR-99 Collection: Description of the WBR-99 Web collection data-structures and file formats. LATIN - Laboratório para o Tratamento de Informação, Dep. de Computação, Universidade Federal de Minas Gerais, Brazil.

(Can & Patton, 2004) Can, Fazli; Patton, Jon M.. "Change of writing style with time." Computers and the Humanities. Vol. 38, No. 1, 2004, p. 61-82. Disponível em http://www.users.muohio.edu/canf/papers/chum04.PDF

(Cardoso *et al*, 2004) Cardoso, Nuno; Silva, Mário J.; Costa, Miguel. The XLDB Group at CLEF 2004. CLEF Workshop in Bath, England. Setembro de 2004. Disponível em http://xldb.fc.ul.pt/data/Publications_attach/cardoso04TheXLDBGroupAtCLEF2004.pdf

(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.frzSzfourelzSzpubzSzfermi96zSzReport-MRIM.pdf /chiaramella96model.pdf

(Chu & Rosenthal, 1996). Chu, Heting; Rosenthal, Marilyn. Search Engines the World Wide Web: A comparative study and evaluation methodology. ASIS 1996. Disponível em 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.

(Cohen, 1995) Cohen, W. 1995. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning.

(Cohen, 1996) Cohen, W. 1996. Learning trees and rules with set-valued *features*. In Fourteenth Conference of the American Association of Artificial Intelligence.

(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. Nº6. 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.

(Cristianini & Shawe-Taylor, 2000) Cristianini, Nello; Shawe-Taylor, John. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 1ª edição, 2000, 189 p.

(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.

(Crystal, 1992) Crystal, D. (1992). An encyclopedic dictionary of language and languages. Oxford: Blackwell Publishers.

(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 Aanalysis. Journal of the American Society for Information Science, 41:391-407.

(Dewdney *et al*, 2001) Dewdney, Nigel; VanEss-Dykema, Carol; McMillan, Richard. The form is the substance: Classification of genres in text. In ACL Workshop on Human Language Technology and Knowledge Management, 2001. Disponível em: www.elsnet.org/km2001/dewdney.pdf

(Dewe *et al*, 1998) Dewe, Johan; Karlgren, Jussi; Bretan, Ivan. Assembling a Balanced *corpus* from the Internet". In Proceedings of the 11th Nordic Conference of Computational Linguistics, Copenhagen, Dinamarca, janeiro de 1998.

(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; Guilloré, Sylvie. 1999. Mutual Expectation: A Measure for Multiword Lexical Unit Extraction. Vextal 99: Venezia per il Trattamento Automatico delle Lingue.

(DiMarco & Hirst, 1990) DiMarco, Chrysanne; Hirst, Graeme. Accounting for style in machine translation. Third International Conference on Theoretical Issues in Machine Translation, Austin, June 1990. Disponível em http://www.mt-archive.info/TMI-1990-DiMarco.pdf .

(DiMarco & Hirst, 1993) DiMarco, Chrysanne; Hirst, Graeme. A Computational Theory of Goal-Directed Style in Syntax. Computational Linguistics.Volume 19, 3, setembro de 1993, p. 451 – 499. Disponível em: http://portal.acm.org/citation.cfm?id=972489

(DiMarco, 1990) DiMarco, Chrysanne. Computational stylistics for natural language translation. PhD thesis, Department of Computer Science, University of Toronto, abril de 1990. Publicada como relatório técnico, CSRI-239.

(Enkvist *et al*, 1974) Enkvist, Nils Erik; Spencer, John; Gregory, Michael J. Lingüística e estilo. Tradução de Wilma A. Assis. 2ª edição. São Paulo, Cultrix, Editora da Universidade de São Paulo, 1974. 126 p.

(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.

(Fidel *et al*, 2004) Fidel, Raya; Pejtersen, Annelise Mark; Cleal, Bryan; Bruce, Harry. A Multidimensional Approach to the Study of Human-Information Interaction: A Case Study of Collaborative Information Retrieval. Journal of the American Society for Information Science and Technology, 55(11):939–953, 2004. Disponível em http://www.ischool.washington.edu/fidelr/RayaPubs/MultiDimensionalApproach.pdf.

(Finn *et al*, 2002) Finn, Aidan; Kushmerick, Nicholas; Smyth, Barry. Genre Classification and Domain Transfer for Information Filtering. Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research, 2002. Disponível em: http://citeseer.ist.psu.edu/finn02genre.html

(Foltz & Dumais, 1992) Foltz, P.W.; Dumais, S.T. Personalized Information Delivery an Analysis of Information Filtering Methods. Communications of the ACM, 35(12), 1992, p. 51–60. Disponível em http://www-psych.nmsu.edu/~pfoltz/cacm/cacm.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.

(Glover & Hirst, 1996) Glover, Angela; Hirst, Graeme. Detecting stylistic inconsistencies in collaborative writing. 1996. In: Sharples, Mike; van der Geest, Thea (editors), The new writing environment: Writers at work in a world of technology. London: Springer-Verlag.

(Glover, 1996) Glover, Angela Doorthy. Automatically detecting stylistic inconsistencies in computer-suported collaborative writting. Tese. Toronto, 1996. Disponível em http://ftp.cs.toronto.edu/pub/gh/Glover-thesis.pdf

(Gomes & Silva, 2005) Gomes, Daniel; Silva, Mário J. Characterizing a National Community Web. ACM Transactions on Internet Technology. Unpublished (in press) 2005. Disponível em: http://xldb.fc.ul.pt/data/Publications_attach/gomesCharacterizing.pdf

(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).

(Gonzalez *et al*, 2005) Gonzalez, Marco; Lima, Vera Lúcia Strube de; Lima, José Valdeni de. Binary Lexical Relations for Text Representation in Information Retrieval. NLDB 2005, p. 21-31.

(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, p. 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, p. 843-857. 2001.

(Gusfield, 1997) Gusfield, D. "Introduction to Suffix Trees" In Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, 1997. Disponível em Webdiis.unizar.es/~jcampos/TAP/material/suffixtrees.pdf

(Gwizdka & Chignell, 1999) Gwizdka, Jacek; Chignell, Mark. Towards Information Retrieval Measures for Evaluation of Web Search Engines. Unpublished manuscript. Disponível em 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. Disponível em 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. Disponível em 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.

(He *et al*, 2004) He, Daqing; Wang, Jianqiang; Luo, Jun; Oard, Douglas W. "iCLEF2004 at Maryland: Summarization Design for Interactive Cross-Language Question Answering", In Carol Peters & Francesca Borri (eds.), Cross Language

Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004). (Bath, Inglaterra, 15-17 de setembro 2004), Pisa, Itália: IST-CNR, p. 267-79.

(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.berkley.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/

(Hirsh & Dinkelacker, 2004) Hirsh, Sandra; Dinkelacker, Jamie. Seeking information in order to produce information: an empirical study at Hewlett Packard labs. Journal of the American Society for Information Science and Technology. Volume 55 , Issue 9 (July 2004), Part II: Information seeking research, Pages: 807 – 817.

(Hosmer & Lemeshow, 2000) Hosmer, David W.; Lemeshow, Stanley. Applied Logistic Regression. Wiley Series in Probability and Statistics - Applied Probability and Statistics. Wiley-Interscience, 2ª edição, 2000, 392 p.

(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.

(Hunt *et al*, 1999) Hunt, Ellen; Jones, Myra; Price, Rebecca; Walker, Claudia; Walker, Lindsay; Williams, Debra. "Register." All American: Literature, History, and Culture. 1999. Disponível em http://www.uncp.edu/home/canada/work/allam/1914-/language/register.htm

(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 & Spink, 2005) Jansen, Bernard J.; Spink, Amanda. An analysis of Web searching by European AlltheWeb.com users. Inf. Process. Manage. 41(2), p. 361-381 2005. Disponível em http://www.cs.helsinki.fi/u/rahholmb/progradu/kallor/Jansen03.pdf

(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, p. 207-227. Disponível em http://jimjansen.tripod.com/academic/acad.html#ResP

(Jansen *et al*, 2005) Jansen, Bernard J.; Spink, Amanda; O. Pedersen, Jan. A temporal comparison of AltaVista Web searching. JASIST 56(6), p. 559-570, 2005.

(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/

(Kalbach, 2003) Kalbach, James. I´m feeling lucky: the role of emotions in seeking information on the Web. Best Practices and Future Visions for Search UIs: A Workshop, CHI 2003, Fort Lauderdale, março de 2003. Disponível em: http://home.earthlink.net/~searchworkshop/docs/JKalbach_Emotions-InformationSeeking-Web_short21.pdf

(Kang & Kim, 2003) Ho Kang, GilChang Kim. Query type classification for Web document retrieval. In Sigir 2003. Toronto, Canada, p. 64 – 71, 2003.

(Karlgren & Cutting, 1994) Karlgren, Jussi; Cutting, Douglass. 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In Proceedings of the 15th International Conference on Computational Linguistics, volume 2, pp. 1071{1075, Kyoto, Japan, August. ICCL.

(Karlgren & Straszheim, 1997) Karlgren, Jussi; Straszheim, Troy. 1997. Visualizing Stylistic Variation". In Proceedings of the 30th Hawaii International Conference on Systems Sciences, Maui, Hawaii, January. IEEE.

(Karlgren *et al*, 1998) Karlgren, Jussi; Bretan, Ivan; Dewe, Johan; Hallberg,Anders; Wolkert, Niklas. 1998. Iterative Information Retrieval Using Fast Clustering and Usage-Specic Genres". In Preben Hansen, editor, Proceedings of Eighth DELOS Workshop on User Interfaces in Digital Libraries, p. 85 - 92, Langholmen. ERCIM.

(Karlgren, 1999) Karlgren, Jussi. Non-topical factors in information access. Invited talk to Webnet´99. Disponível em htpp://www.sics.se/~jussi/papers

(Karlgren, 2000) Karlgren, Jussi. Stylistic Experiments for Information Retrieval. Tese de doutorado, Universidade de Estocolmo, 2000. Disponível em: http://www.sics.se/~jussi/Artiklar/2000_PhD/.

(Karlgren, 2004) Karlgren, Jussi. The wheres and whyfores for studying text genre computationally.In Style and Meaning in Language, Art, Music and Design, Washington D.C., 2004. AAAI Symposium series.

(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.

(Kessler *et al*, 1997) Kessler, Brett; Nunberg, Geoffrey; Schutze, Hinrich. Automatic detection of text genre. In ACL/EACL, 1997. Disponível em: http://acl.ldc.upenn.edu/P/P97/P97-1005.pdf

(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.

(Koppel el al, 2003) Koppel, M.; Akiva, N.; Dagan, I. A *corpus*-Independent *feature* Set for Style Based Text Categorization, in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, 2003. http://www.cs.biu.ac.il/~koppel/papers/stability-workshop.pdf

(Kowalski, 1997) Kowalski, G. Information Retrieval Systems: Theory and Implementation. Boston: Kluwer Academic Publishers, 1997.

(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

(Landwehr *et al*, 2003) Landwehr, N; Hall, M; Frank, E. (2003) Logistic Model Trees. ECML 2003, p. 241-252.

(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 efectiveness 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, 2000) 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 et al, 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/~jgantt/hanspeterluhn.html

(Lyman & Varian, 2000) Peter Lyman, Hal R. Varian. 2000. How Much Information? Extraído de http://www.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf

(Lyman & Varian, 2003) Lyman, Peter; Hal, R. Varian. "How Much Information", 2003. Disponível em http://www.sims.berkeley.edu/how-much-info-2003

(Marcu, 2000) Marcu, D. "The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach." Comp. Ling. 26.3 (2000): 395-448.

(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 & Moreira, 2004) Martins, Junior, J.; Moreira, E. S. Using Support Vector Machines to Recognize Products in E-commerce Pages. In Proceedings of The IASTED International Conference, February 2004, p. 212-217.

(Martins & Silva, 2004) Martins, Bruno; Silva, Mário J. Spelling Correction for Search Engine Queries. EsTAL - España for Natural Language Processing, Alicante, Espanha, outubro de 2004. Disponível em http://xldb.fc.ul.pt/data/Publications_attach/spellcheck.pdf

(Martins *et al*, 2003) Martins, R. T.; Hasegawa, R.; Nunes, M.G.V. Curupira: a functional parser for Brazilian Portuguese. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes (Eds.): Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings. Lecture Notes in Computer Science 2721 Springer 2003, ISBN 3-540-40436-8

(McLachlan, 2004) McLachlan, Geoffrey J. Discriminant Analysis and Statistical Pattern Recognition. Wiley Series in Probability and Statistics. Wiley-Interscience, 2004, 526 p.

(Michos *et al*, 1996) Michos, S. E.; Stamatatos, Fakotakis, N.; Kokkinakis, G. Categorizing Texts By Using a Three Level Functional Style Description , in A. M. Rasmsay (Editor): Artificial Intelligence: Methodology, Systems, Applications, Frontiers in Artificial Intelligence and Applications, Vol. 35, p. 191-198, 105 Press, 1996. Disponível em http://slt.wcl.ee.upatras.gr/papers/michos2.pdf

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

(Mitchell, 2005) Mitchell, Tom M.. Generative and discriminative classifiers: naive bayes and logistic regression. 5 de julho de 2005. Disponível em: http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

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

(Notess, 2000) Notess, Greg R.. Seacrh Engine Statistics: Dead *links* report. Disponível em http://www.notess.com/search/stats/dead.shtml.

(Notess, 2002) Notess, Greg R.. Seacrh Engine Statistics: Unique Hits Report. Disponível em http://www.notess.com/search/stats/unique.html.

(Ntoulas *et al*, 2004) A. Ntoulas; J. Cho; C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In WWW2004. Disponível em http://www.www2004.org/proceedings/docs/contents.htm .

(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

(Orengo & Huyck, 2001) Orengo, V.M.; Huyck, C.R. A Stemming Algorithm for The Portuguese Language. In Proceedings of the SPIRE (String Processing and Information Retrieval) Conference, Laguna de San Raphael, Chile, November 13-15, 2001.

(Orengo & Huyck, 2002) Orengo, V.M.; Huyck, C.R. Portuguese-English Experiments using Latent Semantic Indexing. In Proceedings of the Cross-Language Evaluation Forum (CLEF), Rome 19-20 September, 2002.

(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*, 2003) Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, pp. 210-218 (Lecture

Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.Disponível em: http://www.nilc.icmc.usp.br/~thiago/

(Parris, 1998) Parris, Thomas M. 'Rx for Medical Information On-Line.' Environment, Vol. 40 No. 10, dezembro 1998, p. 3. Disponível em http://environment.harvard.edu/guides/envbon/v40n10.html.

(Payette & Hirst, 1992) Payette, Julie; Hirst, Graeme. An Intelligent computer assistant for stylistic instruction. Computers and the humanities, 26(2), 1992, p. 87-102.

(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. Abril de 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.

(Pizzato & Lima, 2003) Pizzato, Luiz Augusto Sangoi; Lima, Vera Lúcia Strube de. Query Expansion Based on Thesaurus Relations: Evaluation over Internet. CICLing 2003, p. 553-556.

(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.

(Platt, 1998) Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges,; A. Smola, eds., MIT Press. 1998.

(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

(Quinlan, 1992) Quinlan, J. Ross. C4.5 : Programs for Machine Learning. Morgan Kaufmann, 1992, 302 p.

(Quinlan, 1993) Quinlan, Ross. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

(Quirk *et al*, 1992) Quirk, R.; Greenbaum, S.; Leech, G. & Svartvik, J. A A grammar of contemporary English. Longman Group Ltd. Harlow, Reino Unido, 1992.

(Rauber & Müller-Kögler, 2001) Rauber, Andreas; Müller-Kögler, Alexander. Integrating Automatic Genre Analysis into Digital Libraries. ACM/IEEE Joint Conference on Digital Libraries, 2001.

(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 Eletrônicas 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) Robertson, S.E.; Teather, D. A statistical analysis of retrieval tests: a Bayesian approach. Journal of Documentation, 30, p. 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.

(Rose & Levinson, 2004) Rose, Daniel E.; Levinson, Danny. "Understanding user goals in web search", Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 13-19.

(Roussinov *et al*, 2001) Roussinov, Dmitri; Crosswell, Kevin; Nilan, Mike; Kwasnik, Barbara; Cai, Jin; Liu, Xiaoyong. Genre based navigation of the Web. In 34th International Confer-ence on System Sciences, 2001. Disponível em: http://csdl.computer.org/comp/proceedings/hicss/2001/0981/04/09814013.pdf

(Rudman, 2002) Rudman, Joseph: Non-Traditional Authorship Attribution Studies in Eighteenth Century. Stylistics Statistics and the Computer. In: Jahrbuch für Computerphilologie 4 (2002), S. 151-166. Disponível em http://computerphilologie.uni-muenchen.de/jg02/rudman.html#fn79

(Salton & McGill, 1983) Salton, G.; McGill, M. J. (Eds.) (1983). Introduction to Modern Information Retrieval. McGraw-Hill. 448 p.

(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. Regional Conference Series in Applied Mathematics, No. 18, Society of Industrial and Applied Mathematics, Philadelphia, PA.

(Sanderson & Croft, 1999) Sanderson, Mark; Croft, Bruce. Deriving concept hierarchies from text. SIGIR 1999. Disponível em 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 *et al*, 2004) Santos, D.; Simões, A.; Frankenberg-Garcia, A.; Pinto, A.; Barreiro, A.; Maia, B.; Mota, C.; Oliveira, D.; Bick, E.; Ranchhod, E.; Dias de Almeida, J. J.; Cabral, L.; Costa, L.; Sarmento, L.; Chaves, M.; Cardoso, N.; Rocha, P.; Aires, R.; Silva, R.; Vilela, R.; Afonso, S. (2004) Linguateca: Um centro de recursos distribuído para o processamento computacional da língua portuguesa. Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", p. 147-154, IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), novembro de 2004, Puebla - México.

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

(Saracevic & Kantor, 1988) Saracevic, T. & Kantor, P. A study of information seeking and retrieving. III. Searchers, searches and overlap. Journal of the American Society for Information Science, 39, 3, (1988), 197-216. Disponível em http://www.scils.rutgers.edu/~tefko/JASIS1988part3.pdf

(Saracevic, 1995) T. Saracevic. 1995. Evaluation of evaluation in information retrieval. Proceedings og 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.

(Sebastiani, 2002) Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34, 1-47.

(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

(Shepherd, 1997) Adrian J. Shepherd. Second-Order Methods for Neural Networks: Fast and Reliable Training Methods for Multi-Layer Perceptrons. Springer, 1997, 145 p.

(Silva & Oliveira, 2003) Silva, Gilberto; Oliveira, Cláudia. "A Lexicon-Based Stemming Procedure". In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), Computational Processing of the Portuguese Language, 6th International Workshop (PROPOR 2003) (Faro, 26-27 June 2003), Springer Verlag, pp. 159-166.

(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 Review, v. 15, n.6, p. 373-382, 1991.

(Smith, 2002) Lindsay I Smith. A tutorial on Principal Components Analysis. 26 de Fevereiro de 2002. Disponível em: www.cs.otago.ac.nz/cosc453/ student_tutorials/principal_components.pdf

(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 Strzalkovski, editor, Natural Language Information Retrieval. Kluwer, Boston.

(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 .

(Stamatatos *et al*, 1999) E. Stamatatos, N. Fakotakis; G. Kokkinakis. Automatic Authorship Attribution. Eacl 1999. Disponível em http://acl.ldc.upenn.edu/E/E99/E99-1021.pdf

(Stamatatos *et al*, 2000a) E. Stamatatos, N. Fakotakis,; G. Kokkinakis. Text genre detection using common word frequencies. In 18th International Conference on Computational Linguistics, 2000.

(Stamatatos *et al*, 2000b) E. Stamatatos, G. Kokkinakis e N. Fakotakis. Automatic text categorization in terms of genre and author. Computational Linguistics.Volume 26 , Issue 4 (December 2000) Pages: 471 – 495.Disponível em: http://portal.acm.org/citation.cfm?id=971883

(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) Su, L. T.; Chen, H.; Dong, X. 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, p. 348-361.

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

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

(Tombros & Sanderson, 1998) 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. Disponivel em http://ciir.cs.umass.edu/music2000/papers/invites/uitdenbogerd_invite.pdf

(van Rijsbergen, 1979) van Rijsbergen, C. J. 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 Proceding, p. 200. 1986.

(Vieira *et al*, 2000) Vieira, Renata; Gorziza, Fabiano; Rossi, Daniela; Chishman, Rove; Rossoni, Roberta; Pinheiro, 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.

(Whitelaw & Argamon, 2004) Whitelaw, Casey; Argamon, Shlomo. Systemic Functional *features* in Stylistic Text Classification. AAAI Fall Symposim on Style and Meaning in Language, Art, Music, and Design, October 2004. Disponível em http://lingcog.iit.edu/doc/fs804whitelawc.pdf

(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.

(Witten & Frank, 2000) Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann, 2000.

(Wolters & Kirsten, 1999) Maria Wolters; Mathias Kirsten. Exploring the use of linguistic *features* in domain and genre classification. Disponível em http://acl.ldc.upenn.edu/E/E99/E99-1019.pdf

(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. Disponível em 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 quê, quem, quando e onde. 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 significativa com qualquer outra coisa.

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 padrões e sumarizá-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.

Estilística

Em sua definição mais geral, estilística é o estudo de qualquer uso situacional distintivo da língua, e das escolhas feitas por indivíduos e grupos sociais em seus usos da língua (Crystal, 1992, p. 371 apud Glover, 1996). O objetivo da estilística é identificar características estilisticamente significantes da língua (marcadores estilísticos ou de estilo), e as funções de certa forma por elas delimitadas (Crystal, 1992, p. 371 apud Glover, 1996).

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 à 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.

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. Essa 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.

Meta ferramentas de busca

Meta ferramentas de busca ou meta buscadores 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 padrões, só que os procura em textos em língua natural. 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, sendo que a Mineração de Textos 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.

Prefixos

Cadeia de caracteres que inicia uma palavra.

Recuperação de Informação

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 a essa necessidade.

Referenciacão

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.

Registro

É um conceito impreciso utilizado para qualquer variedade da língua associada com diferentes situações e propósitos (Biber & Finegan, 1994). Registros podem ser definidos como dialetos situacionais, tal como o dialeto que usamos quando

falamos com amigos e o dialeto que usamos em uma entrevista de emprego (Hunt et al, 1999).

Resposta automática a perguntas

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

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. Incopora princípios, insights, lições e arquétipos. A sabedoria explica o porquê.

Semântica

Ramo da lingüística que estuda o significado, a relação de significação nos signos e a representação do sentido dos enunciados.

Sintaxe

É a parte da gramática que estuda a disposição das palavras na frase e a das frases no discurso, bem como a relação lógica das frases entre si.

Stopwords

São palavras extremamente freqüentes como artigos, preposições, pronomes e alguns advérbios.

Subcadeias de caracteres

São cadeias de caracteres que podem 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.

Validação cruzada em k partes

Na validação cruzada em k partes, os dados de treinamento são divididos em k subconjuntos de tamanho aproximado, para realizar k treinamentos, cada vez utilizando k-1 para treinamento e 1 para teste.

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 de as páginas dinâmicas possuírem endereços (URLs) que as identificam de forma única, esses não são persistentes.

Apêndice A

Apresentação do Leva-e-traz

Neste apêndice apresenta-se o protótipo de meta buscador Leva-e-traz na forma como foi empregado no experimento com os usuários descrito no Capítulo 10. Foram implementadas no protótipo Leva-e-traz a classificação de resultados de busca na Web segundo necessidades de busca, necessidades personalizadas, gêneros e tipos textuais. Neste apêndice mostramos as telas do Leva-e-traz, inclusive as janelas de ajuda.

Figura 7 – Tela principal do Leva-e-traz

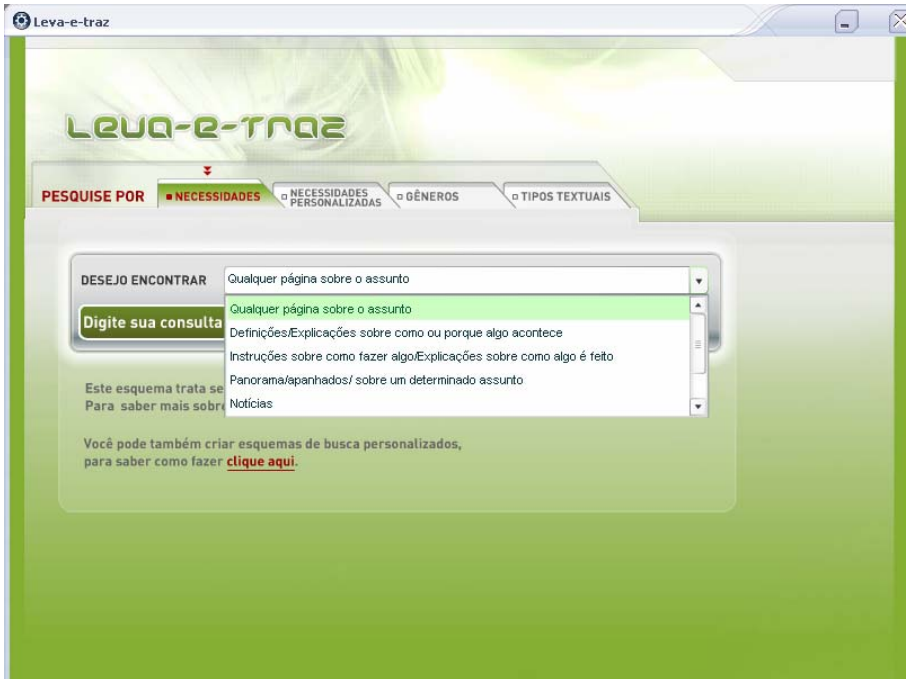

Figura 8 – Escolhendo a opção necessidades

Figura 9 – Escolhendo a opção necessidades personalizadas

Figura 10 – Escolhendo a opção gênero

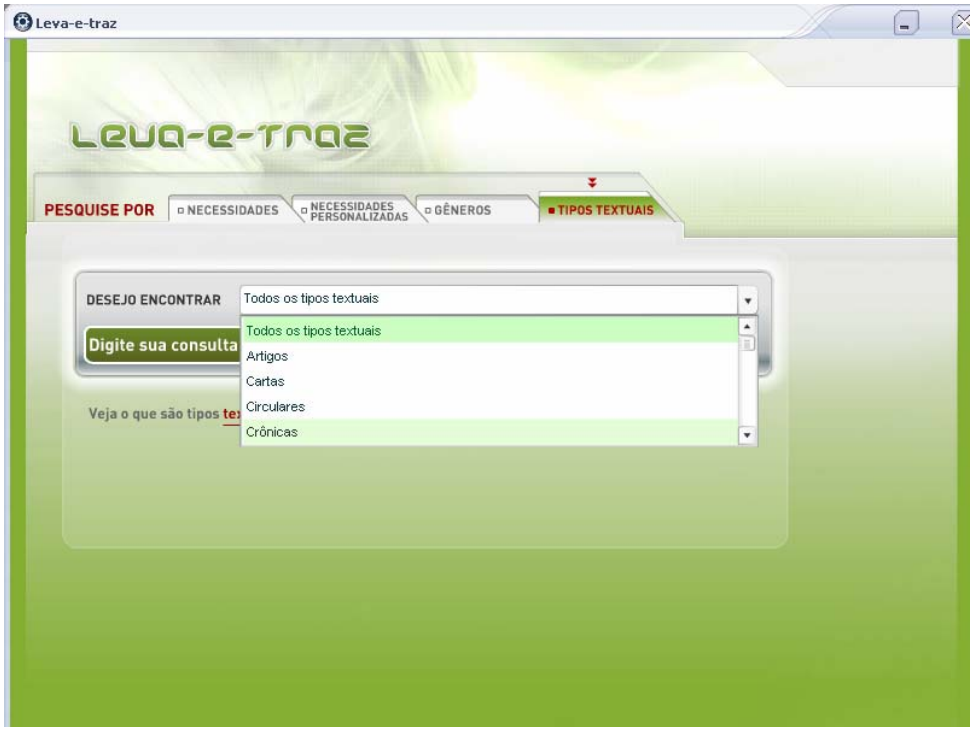

Figura 11 – Escolhendo a opção tipos textuais

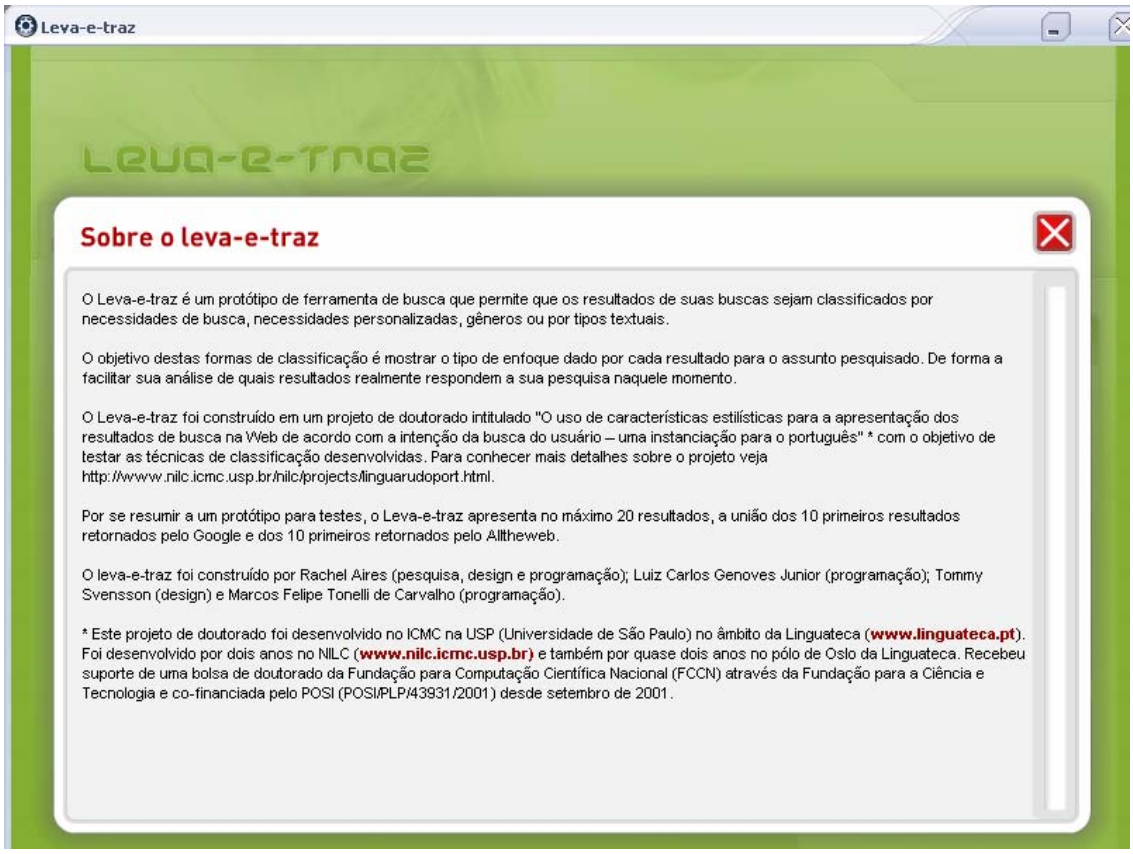

Figura 12– Janela sobre o Leva-e-traz

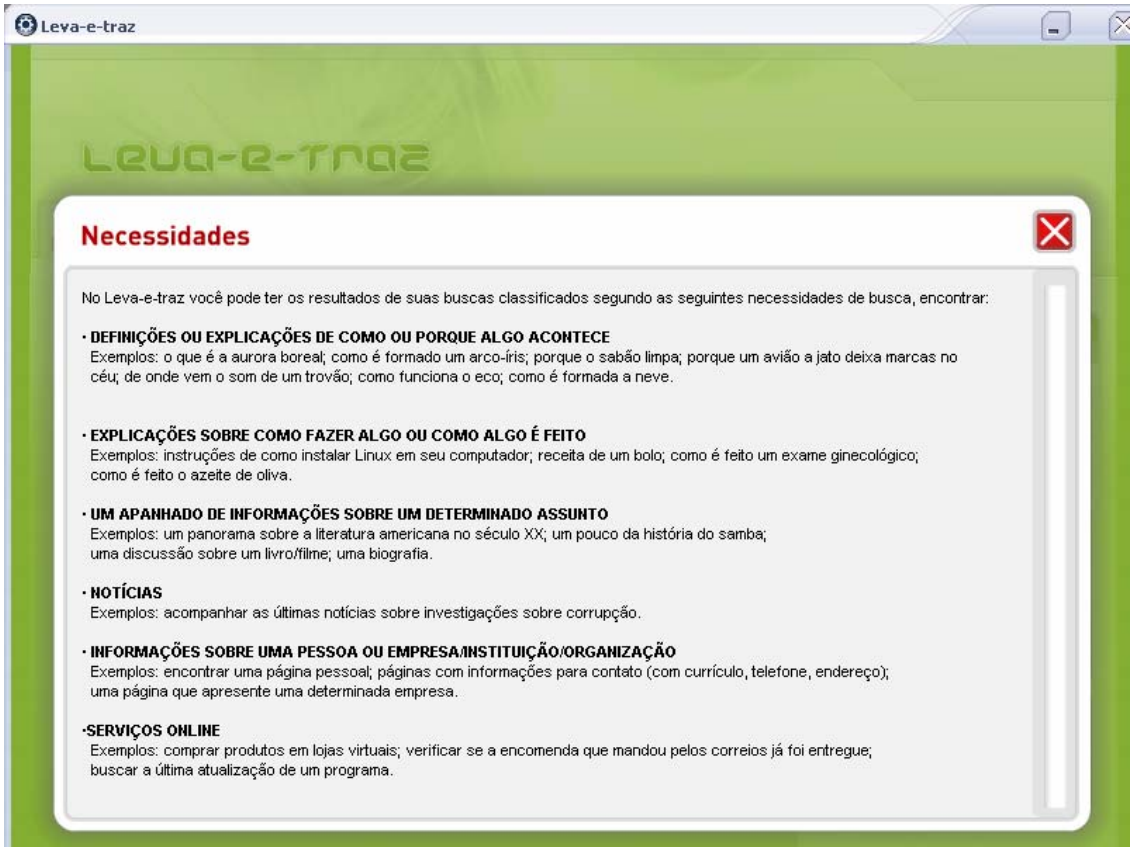

Figura 13 – Janela de ajuda sobre a busca com resultados classificados por necessidades

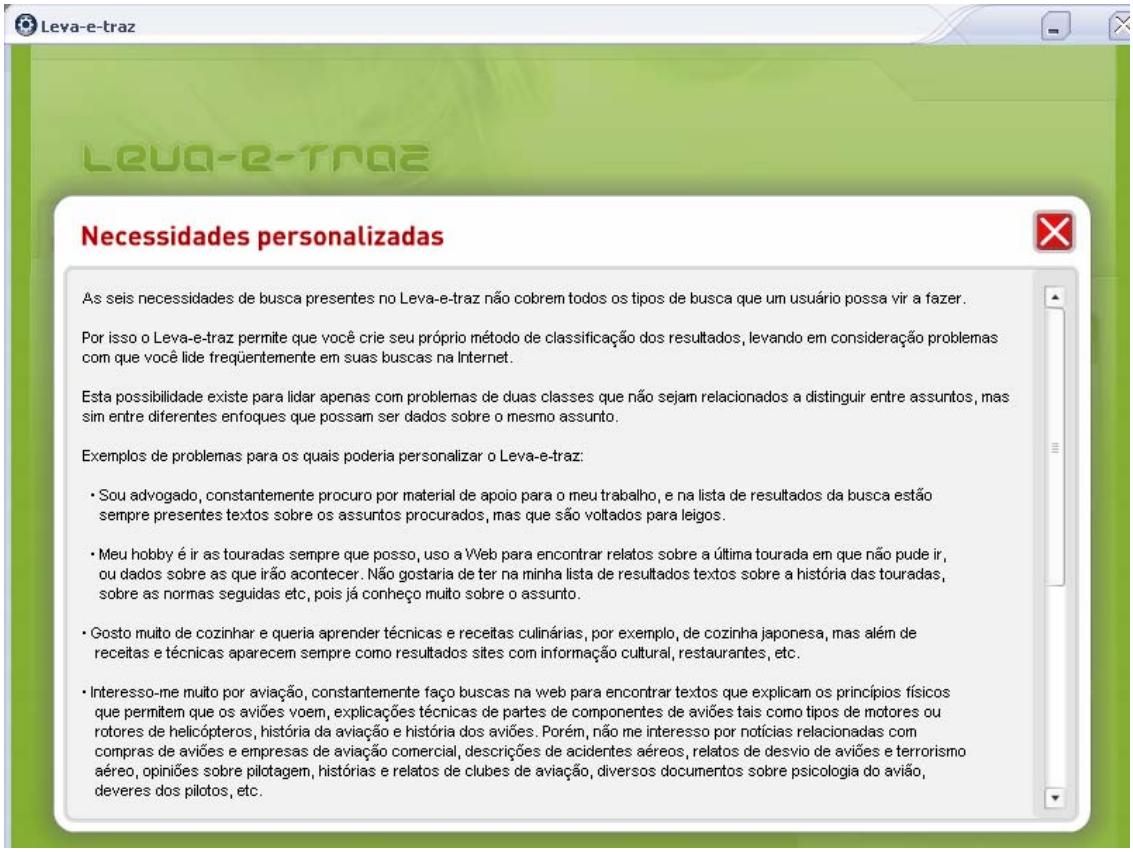

Figura 14 – Janela de ajuda sobre necessidades personalizadas

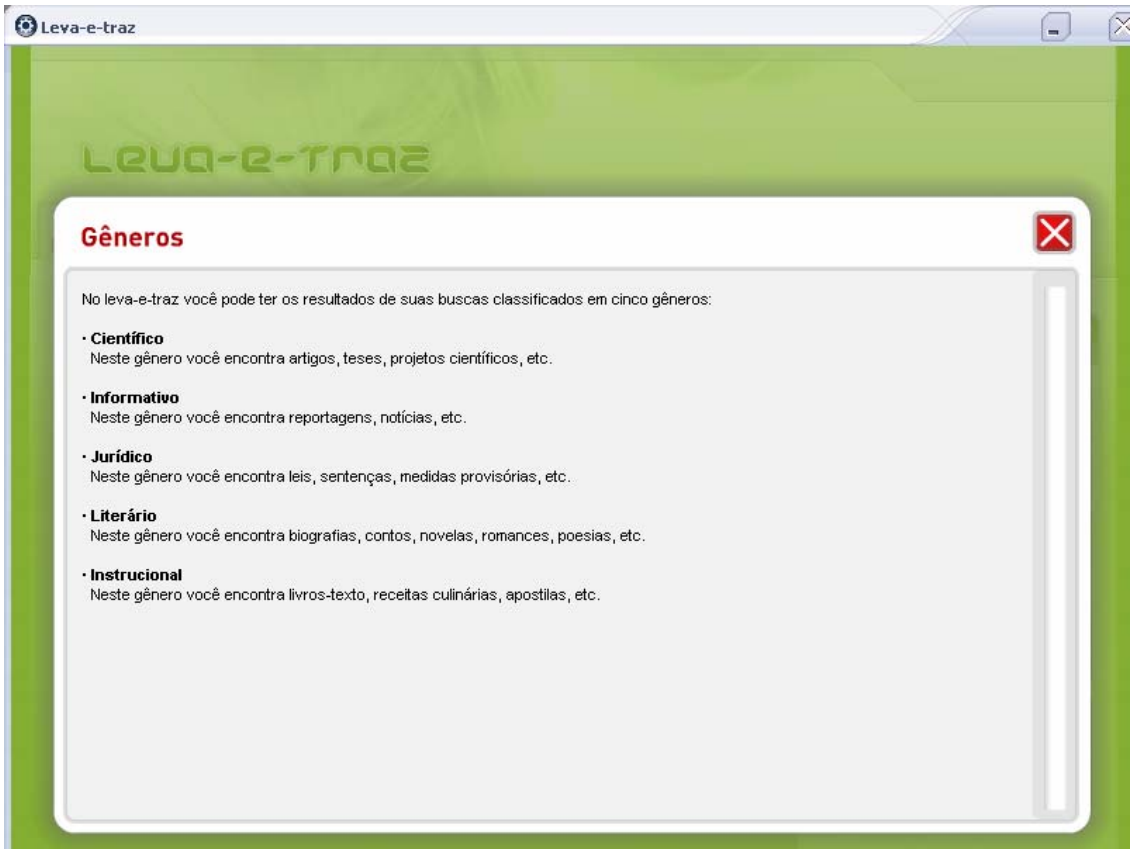

Figura 15 – Janela de ajuda sobre gêneros

Figura 16 – Janela de ajuda sobre tipos textuais

Apêndice B

Questionário inicial

Questionário para dar suporte a uma pesquisa sobre Máquinas de Busca

Este questionário é composto de 4 partes e faz parte de uma pesquisa de doutorado sobre estratégias de melhoria para a busca de textos na Internet, em desenvolvimento no ICMC/USP, cujo objetivo é diminuir ao máximo resultados que não respondem bem a sua consulta. Para conhecer mais detalhes sobre a pesquisa, visite a página:
http://www.nilc.icmc.usp.br/nilc/projects/linguarudoport.html
Agradecemos antecipadamente por sua colaboração.

Parte 1 - Dados pessoais

Instituição: Universidade Federal de São Carlos
Curso: Letras Disciplina: Fonética e Fonologia Semestre:___

Idade:_____os Sexo: ○Feminino ○Masculino

Parte 2 - Experiência em busca na Internet

Locais onde acessa a Internet: ☐ em casa ☐ na faculdade ☐ outros_____

Já utilizou sistemas de busca na Web como p. ex. Google ou Yahoo: ○sim ○não

Você se considera ○inexperiente, ○razoavelmente experiente ou ○muito experiente em efetuar buscas na Internet?

Utiliza serviços de busca: ○raramente ○ocasionalmente/eventualmente ○freqüentemente
Se freqüentemente, qual a freqüência:
_____

Utiliza sistemas de busca para atividades: ☐de lazer ☐do dia a dia ☐ do seu curso ☐ de trabalho
Para quais tarefas usa sistemas de busca com maior freqüência? Por exemplo: encontrar quem vende e quais os preços de produtos; encontrar artigos/trabalhos sobre o trabalho pedido pelo professor.

Nos tipos de busca que mais faz existem tipos de erros que se repetem com freqüência? Quais são? Por exemplo, procurar por uma página que venda um produto e receber como respostas páginas que falam sobre o produto, mas não vendem o mesmo.

Se utilizar sistemas de busca no trabalho:
Qual é a sua profissão/cargo/função?_____
Com que objetivo faz as buscas? Por exemplo: (a) encontrar o programa seguido por outros professores em uma disciplina para planejar o curso que dará no próximo semestre; (b) verificar se o livro que pretende adotar está disponível em várias livrarias; (c) encontrar artigos para se atualizar; (d) pesquisar o valor cobrado por outras pessoas para revisar textos; (e) pesquisar sobre o tema que será abordado em seu próximo livro de ficção.

Você investiria um dia de trabalho para coletar exemplos de textos que seriam utilizados na criação automática de um sistema específico/personalizado para suas necessidades de busca? Por exemplo, para distinguir entre páginas que tragam o resumo de um livro e páginas que tratem da biografia do autor do livro. A coleta de textos seria seu único esforço na criação do sistema, e só teria de ser feita uma vez.

○sim ○não

Cite exemplos de sistemas específicos que você gostaria de ter a sua disposição:

Você estaria disponível para responder mais perguntas sobre sistemas de busca, ou para a avaliação de um novo sistema? Caso a resposta seja sim, qual o seu e-mail?_____

Parte 3 – Busca utilizando tipos de páginas

Imagine um sistema em que cada busca deva ser associada a um tipo de página. As opções são: páginas que
(1) definam alguma coisa ou ensinem como e/ou porque algo acontece;
(2) ensinem como fazer algo ou como algo é feito;
(3) forneçam uma apresentação (ou apanhado ou panorama) sobre um determinado assunto;
(4) apresentem notícias;
(5) forneçam informações sobre uma pessoa ou empresa ou instituição, ou organização;
(6) forneçam algum serviço on-line e
(7) uma página específica que o usuário quer visitar, mas não se lembra da URL.

Qual/quais das sete opções selecionaria para encontrar
- a página oficial do Palmeiras?_____
- uma crítica ou resenha sobre um cd ou show do Skank?_____
- uma lista de hotéis em Araraquara?_____
- a receita de um bolo?_____
- um site para envio de cartões virtuais?_____
- quais as causas do efeito estufa?_____
- como contatar a sede da Embrapa de São Carlos?_____

Acredita que este esquema de classificação de páginas facilitaria suas buscas na Internet? ○sim ○em alguns casos ○precisaria ser melhorado ○não

Tem dúvidas sobre o que algum dos sete tipos abrange?

Acrescentaria outros tipos a estes sete? Uniria alguns dos tipos? Especificaria melhor algum deles?

Parte 4 – Gêneros textuais e seus tipos de texto

Assinale os nomes dos gêneros abaixo que não têm significado totalmente claro para você.

Esquema 1	***Esquema 2***	***Esquema 3***
☐ Científico	☐ Editoriais	☐ Textos privados, informais

☐ Jurídico	☐ Reportagens	☐ Textos públicos, comerciais
☐Técnico-administrativo	☐ Prosa acadêmica	☐ Discussões
☐ De referência	☐ Documentos oficiais	☐ Texto Jornalístico
☐ Instrucional	☐ Literatura	☐ Relatórios, textos técnicos e científicos
☐ Informativo	☐ Receitas	☐ Outros textos
☐ Prosa	☐ Curriculum vitae	☐ Listas e tabelas
☐ Drama	☐ Entrevistas	☐Páginas interativas e formulário
☐ Poesia	☐ Discursos planejados	☐ Coleções de *links*
	☐ Notícias	☐ FAQs
Algum destes três esquemas seria útil para suas buscas na Internet?		
○sim ○não	○sim ○não	○sim ○não

Faça também um X depois do nome de cada gênero acima que possa ser particularmente útil para suas buscas.

Assinale os nomes dos tipos de texto abaixo que não têm significado totalmente claro para você.

☐ artigo, tese, projeto...	☐ livro-texto, receita culinária, apostila,...
☐ lei, sentença, medida provisória,...	☐ biografia, conto, novela, romance..
☐ carta, memorando, manual...	☐ reportagem, notícia, editorial....
☐ enciclopédia, dicionário, glossário,...	

Faça também um X na frente de cada um dos que possam ser particularmente úteis para suas buscas.

Esse esquema de classificação em tipos de texto seria útil para suas buscas na Internet? ○sim ○não

Gostaria de sugerir outros gêneros ou tipos textuais, considerando os tipos textos disponíveis na Internet? Por exemplo, contrato e crônica.

_____
_____
_____
_____
_____
_____
_____

Qual/quais dos cinco esquemas apresentados neste questionário seria mais útil e/ou fácil de utilizar em suas buscas? ☐sete tipos de páginas ☐esquema de gêneros 1 ☐esquema de gêneros 2 ☐esquema de gêneros 3 ☐esquema de tipos de texto

Apêndice C

Questionário final

Questionário sobre busca na Web

Este questionário é composto de 4 partes e faz parte de uma pesquisa de doutorado sobre estratégias de melhoria para a busca de textos na Internet, em desenvolvimento no ICMC/USP. O objetivo da pesquisa é diminuir ao máximo a necessidade de ler resultados de buscas que não respondem bem a uma dada consulta, com o uso de uma interface chamada Leva-e-Traz que organiza os resultados de uma busca. Para conhecer mais detalhes sobre a pesquisa, visite a página:

http://www.nilc.icmc.usp.br/nilc/projects/linguarudoport.html

Na Parte 1 as perguntas são sobre sua experiência de busca; nas Partes 2 e 3 você realizará consultas usando o Leva-e-Traz para atender a 6 tópicos: 3 na Parte 2 e 3 na Parte 3. Na última parte responderá a duas perguntas sobre classificação de textos segundo necessidades de busca.

Agradecemos antecipadamente por sua colaboração.

Parte 1 - Dados pessoais/ Experiência em busca na Internet

Idade:_____ Sexo: ○ Feminino ○ Masculino

Formação:_____

Utiliza serviços de busca como Google, Yahoo e Altavista: ○ raramente

○ ocasionalmente/eventualmente ○ freqüentemente

Você se considera ○ inexperiente, ○ razoavelmente experiente ou ○ muito experiente em efetuar buscas na Internet?

Parte 2 – Três primeiras consultas

Nas consultas para os três tópicos a seguir você deve utilizar a tela principal do Leva-e-traz.

Tópico 1: PREVENÇÃO DA RAIVA EM SERES HUMANOS

O objetivo é encontrar páginas que discutam métodos de prevenção da raiva em pessoas. Uma página para atender a consulta, deve citar pelo menos uma maneira para a prevenção da forma humana da raiva.

Consulta digitada no Leva-e-traz:

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não
6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não

16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente atende melhor a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

Tópico 2: MISÉRIA NA ÁFRICA

O objetivo é encontrar páginas que de tragam informações sobre a miséria na África. Páginas atendem a consulta caso mencionem quaisquer aspectos, tais como causas ou estatísticas sobre a Miséria na África. Notícias sobre eventos para discutir a Miséria na África que não tragam informações sobre a mesma não atendem a consulta.

Consulta digitada no Leva-e-traz:

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não
6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não
16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente atende melhor a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

Tópico 3: MENSALÃO

O objetivo é encontrar páginas que definam o que é e como fazer o recolhimento complementar (mensalão). Notícias ou comentários sobre o escândalo do mensalão não atendem a esta consulta.

Consulta digitada no Leva-e-traz:

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não
6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não
16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente atende melhor a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

Parte 3 – Três consultas utilizando a tab "Necessidades"

Antes de começar leia o texto de help sobre as necessidades. Teve dúvidas? Quais?

_____
_____
_____
_____

Para os três tópicos desta parte você deverá utilizar a tab "Necessidades" do Leva-e-traz.

Tópico 4: O MORRO DOS VENTOS UIVANTES

Encontre páginas que vendam o livro ou o filme O MORRO DOS VENTOS UIVANTES. Páginas que descrevem o livro ou o filme, mas não oferecem os mesmos para venda não atendem a consulta.

Consulta digitada no Leva-e-traz:

_____

Para efetuar sua consulta você escolheu um dos seis tipos de necessidades? Qual?

_____

Teve dúvida ao escolher um dos seis tipos de necessidades? Qual?

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não
6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não
16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Ao julgar um resultado como atendendo ou não a sua consulta, você considerou a classificação da página quanto as necessidades mostrada nas informações do resultado?

○ **sim** ○ **não**

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente melhor atende a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

A classificação te induziu a cliques errados?

○ **sim** ○ **não**

Clicou em resultados corretos que não clicaria não fosse a informação da classificação?

○ **sim** ○ **não**

Tópico 5: BODE E CARNEIRO PARA A PARAPSICOLOGIA

Encontre páginas que definam bode e carneiro segundo a parapsicologia. Páginas que falem de bode e carneiro na parapsicologia, mas não digam o significado destes dois termos não atendem a consulta.

Consulta digitada no Leva-e-traz:

_____

Para efetuar sua consulta você escolheu um dos seis tipos de necessidades? Qual?

_____

Teve dúvida ao escolher um dos seis tipos de necessidades? Qual?

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não

6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não
16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Ao julgar um resultado como atendendo ou não a sua consulta, você considerou a classificação da página quanto as necessidades mostrada nas informações do resultado?

○ **sim** ○ **não**

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente melhor atende a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

A classificação te induziu a cliques errados?

○ **sim** ○ **não**

Clicou em resultados corretos que não clicaria não fosse a informação da classificação?

○ **sim** ○ **não**

Tópico 6: CAUSAS DE INCÊNDIOS DOMÉSTICOS

Seu objetivo é encontrar quais as principais causas de incêndios no lar. Uma página atende a consulta se mencionar pelo menos uma causa de incêndio em residências privadas.

Consulta digitada no Leva-e-traz:

_____

Para efetuar sua consulta você escolheu um dos seis tipos de necessidades? Qual?

_____

Teve dúvida ao escolher um dos seis tipos de necessidades? Qual?

_____

Não clique nos resultados da busca! Leia apenas as informações sobre os resultados disponíveis na tela de resultados, selecione no quadro abaixo quais dos resultados enumerados de 1 a 20 (são retornados no máximo 20 resultados) aparentemente atendem a sua consulta.

1 sim	não	**2** sim	não	**3** sim	não	**4** sim	não	**5** sim	não
6 sim	não	**7** sim	não	**8** sim	não	**9** sim	não	**10** sim	não
11 sim	não	**12** sim	não	**13** sim	não	**14** sim	não	**15** sim	não
16 sim	não	**17** sim	não	**18** sim	não	**19** sim	não	**20** sim	não

Ao julgar um resultado como atendendo ou não a sua consulta, você considerou a classificação da página quanto as necessidades mostrada nas informações do resultado?

○ **sim** ○ **não**

Clique apenas no primeiro resultado da lista. Ele atendia a consulta?

○ **sim** ○ **não**

Dadas as informações disponíveis na tela, **clique no resultado que aparentemente melhor atende a sua consulta**. O resultado selecionado responde a consulta?

○ **sim** ○ **não**

Se seu primeiro clique não respondeu a consulta, **dadas as informações disponíveis na tela continue clicando em resultados de acordo com a sua intuição** de quais respondem a consulta. Quantos cliques foram dados antes de clicar em um resultado que respondesse a consulta?_____

Clique em **todos os resultados para os quais marcou sim no quadro**. Quantos deles atendiam a consulta? _____

A classificação te induziu a cliques errados?

○ **sim** ○ **não**

Clicou em resultados corretos que não clicaria não fosse a informação da classificação?

○ **sim** ○ **não**

Parte 4 – Utilidade da classificação em necessidades

Considera a classificação em necessidades útil?

○ **sim** ○ **não**

Incluiria outros tipos de necessidades de busca? Quais?