Como incrementar a qualidade dos resultados das máquinas de busca: da análise de *logs* à interação em português

Rachel Virgínia Xavier Aires*

Bacharel em ciência da computação - Universidade Católica de Goiás - 1998. Mestre em ciências da computação e matemática computacional – Universidade de São Paulo – 2000. Doutoranda no Programa de Ciências da Computação e Matemática Computacional do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo.
E-mail: raires@icmc.usp.br

Sandra Maria Aluísio

Bacharel em ciência da computação - Universidade Federal de São Carlos - 1986. Mestre em ciências da computação e matemática computacional – Universidade de São Paulo – 1989. Doutora em física computacional – Universidade de São Paulo - 1995.
E-mail: sandra@icmc.usp.br

Resumo

*Com o intuito de avaliar a submissão de consultas em língua natural, especificamente em português, a máquinas de busca na Web, e contrastar com as consultas por palavras-chave, realizou-se um experimento com alunos, professores e funcionários de uma universidade brasileira. Particularmente, analisaram-se as consultas para verificar se os usuários expressavam bem seus objetivos em palavras-chave; como expressariam seus objetivos em língua natural, caso esta possibilidade fosse oferecida; se as consultas em língua natural forneciam informações que pudessem facilitar a recuperação de informação. O pedido de colaboração foi enviado a 440 pessoas de um instituto de computação da universidade. Foram obtidas 63 consultas, correspondentes a 42 objetivos. Observou-se que, para o item **a**, na maioria dos casos (71,43%), as consultas por meio de palavras-chave não trazem todas as informações declaradas importantes no objetivo; para o item **b** as consultas foram feitas por meio de perguntas (71,87%), afirmações (18,75%) e ordens (9,37%); e, para o item **c** todas as perguntas diretas deixavam claro o objetivo da consulta já com a primeira palavra da frase, ou com as duas ou três primeiras, com exceção das iniciadas pela palavra "qual".*

Palavras-chave

*Análise de **logs**; Máquinas de busca; Recuperação de informação; Comportamento de usuários; Estratégias de busca.*

Improving the quality of results of search engines: from Web log analysis to Portuguese natural language interaction

Abstract

*In order to evaluate the use of natural language, specifically Portuguese, in queries of Web search engines, and to contrast it with queries expressed as keywords we carried out an experiment with students and staff of a Brazilian university. Particularly, this experiment was set up to verify: if users convey well their objectives using keywords; b) how users would convey their objectives using natural language if this option was available in search engines; and c) if queries in natural language would improve information retrieval. The data for this experiment were gathered from the response to a consultation via e-mail to students and staff of a Computing Institute of the university (the e-mail message was sent to about 440 people). We gathered 63 queries which corresponded to 42 objectives. The results may be summarized as follows: for item **a**, most queries (71,43%), using keywords do not bring all the information regarded as important; for item **b**, the queries were conveyed as questions (71,87%), statements (18,75%) and orders (9,37%); and for item **c**, all the direct questions made explicit the objective using only the first word of the sentence or using two or three words, except in the case of which-type questions.*

Keywords

Log analysis; Search engines; Information retrieval; Users´ behavior; Search strategies.

* Projeto Processamento Computacional do Português – Sintef Telecom and Informatics. Pb 124, Blindern - NO-0314 Oslo – Noruega.

INTRODUÇÃO

A propagação da Internet propiciou o estudo de vários tópicos relacionados ao uso da *Web*. Porém, ainda são poucos os estudos sobre o comportamento e padrões de perguntas dos usuários utilizadores de máquinas de busca (*search engines*) ou de diretórios (*Internet directory*) (Jansen & Pooch; 2000, Cacheda & Viña, 2001a), apesar de 85% dos usuários da *Web* utilizarem estes serviços (Kobayashi & Takeda, 2000). Entre os pioneiros estão algumas estatísticas sobre buscas via Infoseek apresentadas por Kirsch (1998), os estudos de Jansen *et alii* (1998) sobre consultas feitas no Excite e os de Silverstein *et alii* (1999) sobre consultas realizadas no Altavista.

Já os sistemas de recuperação de informação (RI) tradicionais e catálogos de acesso público têm sido estudados há mais de 40 anos (Jansen & Pooch, 2000). Jansen & Pooch (2000) levantam como possível razão o fato de ser extremamente difícil realizar estudos válidos sobre buscas em um ambiente como a *Web* e sugerem um *framework* para futuros estudos. Os autores acreditam que, utilizando um *framework* comum, será possível comparar os resultados entre diferentes estudos, possibilitando responder a questões como: a) *todas as características de busca levantadas são comuns para quaisquer usuários deste tipo de sistema?*; b) a*s características são conseqüência de algum aspecto do sistema ou da coleção de documentos?*; c) o*s usuários que participaram dos estudos fazem parte da mesma população ou são de populações diferentes?*

Se o número de estudos já é limitado quando se trata da análise de consultas submetidas a sistemas de busca genéricos na *Web* (aqueles que aceitam perguntas em várias línguas) ou quando se trata de sistemas de busca em inglês, o número de estudos para sistemas de busca de países cujo idioma oficial não é o inglês é praticamente nulo. Em nossa pesquisa, encontramos apenas três estudos sobre consultas submetidas a sistemas de busca na *Web* de países cujo idioma oficial não é o inglês, dois sobre os *transaction logs* (veja na próxima seção) de um diretório da Internet espanhola chamado Biwe* (Internet Searcher of Spanish Webs) (Cacheda & Viña, 2001a, 2001b) e um que inclui consultas feitas a um servidor *Web* coreano (Abdulla *et alii*, 1997). Não encontramos

* http://www.biwe.es (20/04/2003)

estudos que analisassem consultas submetidas a sistemas de busca na *Web* brasileira, portuguesa ou de qualquer um dos cinco países africanos cuja língua oficial é o português. Em se tratando de sistemas de busca em países cujo idioma oficial é o português, só encontramos trabalhos descrevendo sistemas, analisando seu desempenho ou comparando seu desempenho com outros sistemas. Entretanto, o português é usado diariamente por cerca de 220 milhões de pessoas. O português é língua oficial no Brasil, Portugal, Angola, Cabo Verde, Guiné Bissau, Moçambique, S. Tomé e Príncipe e, também, no Timor Leste. Depois do inglês e do africano, línguas oficiais na África do Sul, o português surge como primeira língua estrangeira com 1 milhão de usuários. Devido à influência portuguesa ou à proximidade com Angola e Moçambique, existem comunidades que falam português no Senegal, Zâmbia e no Zimbábue. Na Namíbia, fronteira com Angola, um em cada cinco habitantes é falante do português. Em Macau, na China, o português é língua oficial. Em Goa, Damão e Diu, na Índia, há um grupo significante de usuários do português, e há outro grupo até mesmo no Vietnã. O português é ensinado em 29 universidades japonesas, e em quatro delas existem cursos de mestrado no idioma. Não podemos esquecer a emigração de brasileiros, portugueses e africanos que começam a formar comunidades de usuários do português nos EUA.

Sabemos que apenas parte destes usuários do idioma português tem acesso à Internet, mas esta parte fez com que o português fosse a quarta língua mais utilizada na Internet. De qualquer forma, já em 1999, o Brasil tinha mais usuários com acesso à Internet do que a França, a Austrália e a China* (as estatísticas mostradas no *site* NUA** apontam 11,94 milhões de usuários no Brasil até julho de 2001 e 3,6 milhões em Portugal até dezembro de 2001***). A perspectiva é que, em 2003, 36 milhões de brasileiros terão acesso a Internet**** . Tal expectativa é baseada na taxa de crescimento anual do número de usuários da Internet no Brasil e também no programa do governo brasileiro de Internet pública, chamado Projeto Sociedade da Informação. O projeto prevê a implantação da Internet em todas as escolas e bibliotecas públicas, em museus e em 5 mil ONGs até 2003. Outra idéia do projeto é aproveitar a infra-estrutura dos correios para criar postos onde os usuários possam ter seu endereço eletrônico. Há também planos de implantação em bancas de jornal, lotéricas e lojas de conveniência*. Não podemos nos esquecer também dos celulares – 29.545.524, em abril de 2002, segundo dados da Anatel**.

O experimento retratado neste artigo faz parte de um estudo maior, cujo objetivo é delimitar uma abordagem para RI Textual para português que utilize Processamento de Língua Natural (PLN) em todas as fases do processo de RI e aplicar esta abordagem na tarefa de busca. Neste experimento, analisamos consultas em português feitas a máquinas de busca para verificar: a) se os usuários expressavam bem seus objetivos em palavras-chave; b) como os usuários expressariam seus objetivos em língua natural, caso esta possibilidade fosse oferecida pelas máquinas de busca; c) se as consultas em língua natural forneciam de alguma maneira informações que pudessem facilitar a recuperação de informação.

AVALIAÇÃO DO COMPORTAMENTO DOS USUÁRIOS DE SISTEMAS DE BUSCA

O instrumento utilizado para as avaliações do comportamento dos usuários de sistemas de busca na Internet tem sido os *transaction logs*. Apesar de suas limitações por só lidar com as ações dos usuários, este método comumente utilizado para capturar características das interações de usuários com sistemas de RI (Peters, 1993) tem sido utilizado por ser uma forma não intrusiva de coletar informações sobre buscas.

As informações contidas em um *transaction log* variam de um estudo de sistemas de busca para outro. Jansen & Pooch (2000) apresentam um bom resumo das características de vários estudos analisados. As análises podem ser feitas em três níveis – sessões, consultas e termos das consultas – e são detalhadas a seguir.

As sessões compreendem o intervalo em que um determinado usuário esteve utilizando um sistema de busca. Os estudos que levam em conta as sessões consideram cada IP como um usuário. No entanto, é possível que diferentes usuários compartilhem o mesmo computador. Uma alternativa é delimitar um tempo máximo para uma sessão (Silverstein *et alii*, 1999). As análises no nível de sessão são úteis, por exemplo,

* www.estadao.estadao.com.br/colunistas/zero/2000/07/zero000706.html (10/09/2002)

** http://www.nua.ie/surveys/how_many_online/s_america.html (20/04/2003)

*** http://www.nua.ie/surveys/how_many_online/europe.html (20/04/2003)

**** http://www.rnw.nl/parceria/html/internet.html (20/04/2003)

* http://www2.globo.com/infotech/arquivo/materias/20010507/4whpmv.htm (20/04/2003)

** http://200.252.158.174/srem_smc/srem_smc_consulta/geralanatelevolucao.asp (10/09/2002)

para determinar quanto tempo em média um usuário leva em uma busca, quantas páginas este usuário visita durante este intervalo de tempo e quantas consultas um usuário faz por sessão. Assim como nos outros níveis, há decisões que podem variar de um estudo para outro (Jansen & Pooch, 2000). Por exemplo, "se um usuário acessa a página do sistema de busca e não digita nada, ainda assim o tempo da sessão começa a ser contado?"; "se o sistema gera opções de consultas para mostrar os resultados, como faz, por exemplo, o Ask Jeeves (http://www.askjeeves.com), estas consultas farão parte da sessão?".

As consultas são constituídas de todos os caracteres como foram digitados por um usuário para realizar uma busca; incluem palavras e operadores que tenham sido utilizados. Neste nível de análise, são levantadas estatísticas sobre o número de usuários que utilizam operadores booleanos, que utilizam características avançadas do sistema de busca e que fazem as consultas por meio de frases e, também, sobre o número de termos por consulta, se há usuários fazendo buscas de forma que o sistema não previa, entre outras.

Os termos são as cadeias da consulta separadas por algum delimitador. Uma opção é não considerar como termos os operadores, por exemplo: "and", "or", "+", "–". Uma vantagem óbvia de remover os operadores é que os operadores incluídos pelo sistema na criação de sua versão interna da consulta não farão parte da análise. Mas, como saber se é um operador ou uma conjunção? Análises neste nível podem ser utilizadas para gerar listas dos assuntos mais procurados e calcular o número médio de palavras por consulta, por exemplo.

Estes níveis de análise são interpretados em separado ou em conjunto e possibilitam, entre outras coisas, saber:

• se os usuários repetem uma consulta. Existem formas diversas de caracterizar uma consulta como diferente da anterior: pela ordem das palavras, pelo número de palavras, pelo tamanho da consulta, entre outras;

• quantos usuários visitam a mesma página mais de uma vez. Saber quantos usuários diferentes visitam uma determinada página e em que intervalo de tempo. Isto pode ser útil para definir como guardar páginas em *cache*, por quanto tempo guardar, que páginas guardar (Markatos, 2000);

• que funções de busca, busca avançada e operadores os usuários utilizam. Isto permite saber o que colocar em evidência na página principal, o que pode ficar mais escondido;

• quais são os temas mais procurados, quais são as áreas de interesse dos usuários. Os diretórios, por exemplo, podem dar mais atenção na busca de documentos de um dado assunto ou até utilizar as áreas de interesse para reorganizar as categorias ou criar novas. Outro fato é que os usuários gostam de saber o que as outras pessoas estão procurando (Jansen & Pooch, 2000). Assim, os temas mais procurados podem dar origem a páginas como uma chamada "Mais procuradas de 2001" (http://buscadores.hpg.ig.com.br/palavras.htm), que diz que a beleza masculina ou a fantasia sexual via *Web* não seduzem tanto as mulheres brasileiras quanto seduzem os homens. Ou as listas divulgadas por sistemas como o AskJeeves e Metacrawler (http://www.metacrawler.com);

• o número de páginas (telas) de resultados que os usuários vêem. Este número é utilizado para justificar a importância de se ter um bom algoritmo de *ranking* (Cacheda & Viña, 2001a, 2001b; Silverstein *et alii,* 1999; Jansen *et alii,* 1998);

• o número de palavras por consulta. Os artigos que fazem este cálculo não dizem para que o utilizam ou utilizariam;

• com que freqüência os usuários modificam uma consulta. Analisam também como as consultas foram alteradas, se acrescentando ou subtraindo termos ou se apenas trocando um termo por outro (Spink *et alii*, 2001; Wolfram *et alii*, 2001);

• o número de resultados para uma dada consulta. Esta informação pode ser utilizada para saber quantos usuários fazem consultas muito genéricas, imaginando-se que, se uma consulta é muito grande, esta consulta é vaga (e pode ser ambígua) (Cacheda & Viña, 2001b).

• o número de *links* visitados. Esta informação pode sugerir dar uma importância maior à criação de resumos sobre documentos, caso se verifique, como em Cacheda e Viña (2001a), que apenas três documentos são acessados por sessão, o que sugere para os autores que usuários só abrem documentos cujos títulos e descrições descrevem melhor o que eles procuram;

• o número de usuários que utilizam as consultas modificadas geradas pelos sistemas, como é feito no AskJeeves;

• o número de usuários que utilizam as alternativas dadas por sistemas para a forma correta de escrever uma palavra, como faz o Google (http://www.google.com.br).

• o número de usuários que utilizam o *feedback de relevância* como a opção "*More like this*" do Excite (Spink *et alii*, 2001).

Os artigos de Cacheda & Viña (2001a, 2001b), Jansen & Spink (2000), Jansen *et alii* (1998), Jansen, Spink & Saracevic (2000), Ross & Wolfrmam (2000), Silversten *et alii* (1998, 1999), Spink *et alii* (2001) e Spink & Xu (2000) são exemplos de estudos em um ou mais dos três níveis de análise. As estatísticas e os assuntos mais procurados (assuntos, e não categorias) variam de um estudo para outro, mas corroboram a idéia de que os usuários fazem pesquisas simples: fazem poucas consultas; formulam consultas com poucos termos; poucos são os que acrescentam ou removem termos ao tentar novamente uma consulta; em geral, apenas trocam um termo por outro; checam poucas páginas de resultados; visitam poucos *links*; e poucos utilizam técnicas avançadas de busca, operadores booleanos e o *feedback de relevância*.

Estudos longitudinais foram feitos pela equipe que analisa as consultas do Excite. No artigo de Wolfram *et alii* (2001), são comparadas as estatísticas das consultas coletadas em 1997 com as coletadas em 1999, e no artigo de Spink *et alii* (2002) são comparados os resultados de 1997, 1999 e 2001. As comparações mostram que o comportamento do usuário ao fazer as buscas não sofreu mudanças significativas, que apenas a ordem das 11 categorias mais procuradas de informação tem mudado. Em 1997, o número 1 da lista era a categoria *entertainment or recreation* com 19,9% das buscas, em 1999 e 2001 o número 1 passou a ser a categoria *commerce, travel, employment or economy*, com respectivamente 24,5% e 24,7% das buscas (Spink *et alii*, 2002). As três listas demonstram um amadurecimento das necessidades de busca que coincide com as mudanças da distribuição da informação nas páginas publicamente indexadas da *Web*. Por exemplo, em 1999, 83% dos servidores *Web* continham conteúdo comercial (Lawrence & Giles, 1999). A *Web* tem se tornado mais complexa, e as necessidades dos usuários, também. Por exemplo, a maioria das pessoas que procura por sexo (quinta categoria da lista em 2001) procura por informações mais relacionadas à saúde e sexualidade, do que a pornografia (Wolfram *et alii*, 2001). Outro ponto que merece destaque é a mudança de posição da categoria *non english or unknown*, 10º lugar em 1997 com 4,1%, 7º em 1999 com 6,8% e 3º em 2001 com 11,3%, podendo indicar que outros idiomas diferentes do inglês estão conquistando mais espaços na *Web* e também que a *Web* nestes idiomas está se desenvolvendo rapidamente. Esta categoria inclui palavras que não são da língua inglesa, como, por exemplo, números e siglas.

Em 1997, as consultas eram simples, em 1999 as consultas continuavam simples e o mesmo aconteceu em 2001. O que isto implica? É esta a razão de continuarmos insatisfeitos com os sistemas de busca na *Web*, isto é, a má elaboração das consultas? Jansen (2000) investigou o efeito da estrutura das consultas nos resultados recuperados por sistemas de busca na *Web*. Selecionou 15 consultas simples (sem qualquer recurso avançado, como o uso de operadores booleanos ou pesquisa por frase exata) de um *transation log* e as submeteu a cinco máquinas de busca: Alta Vista (www.altavista.com/), Excite (www.excite.com/), FAST Search (http://www.alltheweb.com/), Infoseek (infoseek.go.com/) e Northern Light (www.northernlight.com/). As 15 consultas foram modificadas utilizando os operadores suportados em cada uma das cinco máquinas de busca, o que deu origem a 210 consultas complexas que foram submetidas a cada uma das máquinas de busca. Comparando os resultados, Jansen conclui que utilizar consultas complexas aparentemente tem um impacto muito pequeno nos resultados recuperados, tais consultas em média trazem apenas 2,7 resultados diferentes dos que haviam sido apresentados pelas consultas simples.

Após analisar todos estes estudos, concordamos com Spink *et alii* (2002), quando dizem que precisamos de uma nova geração de máquina de busca que seja baseada em maior entendimento do comportamento de busca de informação de humanos.

AVALIAÇÃO DE *LOGS* EM PORTUGUÊS

Além das restrições relacionadas ao uso de *transaction logs*, outra questão a ser levantada sobre os estudos anteriormente comentados é o fato de eles analisarem a população de um único sistema de busca. O grupo que faz seus estudos sobre dados do Excite assume que usar dados apenas do Excite é uma limitação em seu estudo, e no mesmo artigo é dito que estão expandindo seu estudo com a análise de *logs* de consultas do Fast.no (Spink *et alii*, 2002). Nossa pergunta é: "Todos os resultados encontrados até agora são válidos para os usuários do idioma português?".

Fora diferenças culturais que existem até mesmo entre Brasil, Portugal e países da África, há também a questão de a Internet não ter crescido com a mesma proporção simultaneamente nestes países e nos Estados Unidos (os estudos dos quais tratamos são baseados em *logs* de sistemas de busca americanos).

Temos várias suposições sobre pontos em comum e diferenças entre os usuários do português e os demais. Acreditamos, por nossas experiências pessoais, que boa parte dos usuários do português não vê mais do que a primeira página dos resultados e que as sessões devem mesmo ter cerca de 10 minutos de duração. Imaginamos que, devido a características do português, o número de termos por consulta possa ser diferente da média do inglês, mesmo que nossos usuários também tendam a fazer consultas simples. É possível supor que nossos usuários irão utilizar "e" e "não" no lugar dos operadores *and* e *not*, assim como os usuários do espanhol utilizam *y* e *no*. Podemos concordar com Cacheda & Viña (2001a) em que os sistemas de busca devem interpretar os operadores lógicos de seu idioma de origem. E, finalmente, podemos divagar sobre as diferenças entre os temas que interessam nossos usuários e os usuários do inglês. Diferenças originadas pela diferença do estado da *Web* para as duas línguas (conteúdo) e pela diferença de hábito dos usuários com estes sistemas, como também originadas pelas diferenças culturais. Podemos dizer que futebol é um dos temas mais cotados no Brasil, que, em 2002, um dos temas mais cotados será a "Copa do Mundo" e que este não será um tema tão cotado para usuários noruegueses como "Olimpíadas de inverno", principalmente porque em 2002 a seleção da Noruega não está na Copa. No entanto, no papel de cientistas, e não de adivinhos, preferimos a análise de dados a dar nossa opinião.

Um fato que entrava a análise de *logs*, tanto em inglês como em português, é que máquinas de busca comerciais tradicionalmente não fornecem seus *logs*. O artigo de Spink *et alii* (2002) nos lembra que obter grandes *logs* de máquinas de busca comerciais não é uma tarefa fácil. E realmente não é. Solicitamos a 19 sistemas de busca da *Web* os *logs* das consultas de seus usuários (perguntas somente) e nos comprometemos a fornecer todos os resultados de nossa análise a quem nos fornecesse os dados. Dos 13 sistemas brasileiros e 6 portugueses, apenas um se dispôs a nos ajudar, mas não fornecendo os *logs*, e sim apenas algumas estatísticas. Dadas estas circunstâncias, resolvemos investigar parte do que nos interessava com dados obtidos diretamente com usuários de uma universidade. Partindo do pressuposto de que uma máquina de busca que possibilita a interação em língua natural é uma alternativa de interface muito mais interessante para qualquer usuário humano, pretendíamos investigar o que estes usuários poderiam fornecer em uma interação natural que poderia ser aproveitado pelas máquinas de busca para retornar resultados mais precisos.

ANÁLISE DE *LOGS* REALIZADA

O experimento

Os dados para este experimento foram coletados mediante um pedido feito por *e-mail* a cerca de 440 pessoas (estudantes, professores e funcionários) de um instituto de computação de uma universidade brasileira. Pedimos que armazenassem em um arquivo as consultas que realizassem em português no período de um mês, acompanhadas dos seus objetivos e das formas correspondentes em língua natural. O fato de ter o objetivo explicitado era importante para que não restassem dúvidas sobre o que o usuário pretendia encontrar com cada consulta, possibilitando analisar se as consultas por meio de palavras-chave e as em língua natural forneciam as informações necessárias ou se eram de alguma forma incompletas ou ambíguas.

É importante ressaltar que, apesar de ser também de nosso interesse analisar as máquinas de busca para português, não fazia parte deste experimento avaliar o desempenho das máquinas de busca ou a satisfação do usuário com relação às mesmas. Por isso, não pedimos que os usuários identificassem as máquinas de busca utilizadas, nem dissessem se estavam satisfeitos com os resultados de cada consulta ou fornecessem os resultados obtidos.

RESULTADOS

Obtivemos 63 consultas (correspondentes a 42 objetivos) em resposta à solicitação feita por *e-mail*. Parte das consultas é mostrada na tabela 1. Das cerca de 440 pessoas, apenas 16 enviaram suas consultas. Destas 16 pessoas, apenas 11 enviaram as consultas acompanhadas de objetivo e consulta em língua natural, por isso 12 consultas recebidas foram desconsideradas nesta análise e não estão entre as 63 listadas na tabela 1, a seguir. Oitenta por cento dos colaboradores são pessoas da área de computação, e 20%, da área de letras. Oitenta por cento são estudantes de pós-graduação, e 20%, estudantes de graduação.

TABELA 1
Consultas recebidas

	Consulta por meio de palavras-chave	Objetivo	Consulta em língua natural
01	RegClean	Encontrar o programa RegClean para limpar registros do Windows	Onde eu posso encontrar o programa RegClean
02	"Cd writer" +"driver"	Encontrar o driver do cd writer da HP	Onde encontrar o driver do cd writer da HP
03	"Algoritmo EM" +"estimação"	Encontrar documentação sobre o algoritomo EM	Onde posso encontrar documentação sobre o algoritmo EM
04	Biederman	Encontrar um artigo de Biederman sobre reconhecimento por componentes (Recognition by Components)	Artigo Recognition by Components escrito por Biederman
05	Lendas brasileiras	Encontrar um site com lendas brasileiras	Quero um site de lendas brasileiras
06	Site lendas brasileiras		
07	Content Word	Encontrar a definição de content word e a denominação equivalente em português	O que é content word
08	Regressão Linear Múltipla	Encontrar uma explicação do que é Regressão Linear Múltipla, para que serve, e exemplos ilustrativos	O que é Regressão Linear Múltipla
09	Richard Bellman Dynamic	Encontrar páginas sobre Dynamic Programming, técnica cujo autor é o Bellman	O que é Dynamic Programming
10	SIMR	Encontrar aplicações para o algoritmo SIMR de reconhecimento de padrão	Em quais aplicações o SIMR é usado
11	Sooth Injective Map Recognizer		
12	Syntactic pattern recognition	Encontrar páginas sobre reconhecimento sintático de padrão, principalmente com uma definição e as aplicações mais relevantes	O que é e para que serve reconhecimento sintático de padrão
13	Borland	Encontrar a homepage da borland	Qual é a homepage da borland em português
14	Uml diagrama de estados	Encontrar textos sobre análise orientada a objetos com Uml	Como fazer análise orientada a objetos com Uml
15	Salmo101	Encontrar uma página com o salmo 101	Como é o salmo 101
16	+delphi +componente +matemática	Encontrar um componente do Delphi para avaliar e computar funções matemáticas	"Componentes do Delphi para funções matemáticas"
17	"Ícones para páginas"	Encontrar arquivo de figura (gif ou jpg) de página em construção	Onde posso encontrar um ícone de página em construção
18	"Figuras para páginas"		
19	"Dependência conceitual"	Encontrar trabalhos em português sobre a teoria de Dependência Conceitual de Schank	Como é a teoria de dependência conceitual de Schank
20	+delphi +componente + "campo numérico"	Encontrar um componente de edição do Delphi para permitir somente a entrada de valores numéricos	"componentes do Delphi para campos numéricos"

	Consulta por meio de palavras-chave	Objetivo	Consulta em língua natural
21	"O Morro dos Ventos Uivantes"	Estava procurando sites que falassem do livro ou do filme "O Morro dos Ventos Uivantes" de Emily Bronte	Onde posso encontrar sites sobre o livro ou o filme "O Morro dos Ventos Uivantes", com ilustrações e críticas, em português
22	"Morro*Ventos*Uivantes"		
23	'Morro*Uivantes"		
24	"Livrarias"	Estava procurando sites de livrarias on-line, em São Paulo, que fizessem entregas	Onde posso encontrar sites de livrarias em São Paulo, com serviços de entrega e com livros relacionados à literatura e educação
25	"Livrarias on-line"		
26	"Livros"	Estava procurando resumos de livros disponíveis na Internet	Onde posso encontrar resumos de livros, em especial para Vestibular
27	"Resumos de Livros"		
28	"Anjos"	Estava procurando imagens de anjos	Onde posso encontrar belas figuras de anjos
29	"Cinderela"	Estava procurando sites que possuíssem fotos da história da Cinderela para festas infantis	Onde posso encontrar exemplos de decoração para festas infantis, com ilustrações da história da Cinderela
30	"Decoração de festas"		
31	Loteria federal	Saber o resultado do sorteio da loteria federal de 27/10/01	Qual foi o resultado da loteria federal de hoje
32	Beatles lyrics	Encontrar a letra da música The Ballad of John e Yoko	Como é a letra da música The Ballad of John and Yoko
33	John Swales	Encontrar publicações de John Swales	Quais são as publicações de John Swales
34	Crônicas e crônica	Encontrar sites que falassem sobre o gênero literário "crônica" e que trouxessem crônicas, incluindo sites de jornais e revistas	Quero uma relação de crônicas e/ou cronistas O que é a crônica
35	"Algoritmo para números primos"	Estava tentando encontrar um algoritmo específico que implementasse seleção de números primos dado em sala de aula, mas não me lembrava o nome	Algoritmos para seleção de números primos
36	"Algoritmo para multiplicação de matrizes"	Estava tentando encontrar um algoritmo específico que implementasse multiplicação de matrizes dado em sala de aula, mas não me lembrava o nome	Algoritmos para mulitplicação de matrizes
37	"Dia and windows"	Estava tentando encontrar uma versão do software ´dia´ (semelhante ao Visio) para windows	Uma versão do software dia para windows
38	"Driver SiS5595"	Encontrar driver de instalação da placa de vídeo SiS5595 (silicon)	Driver de instalação da placa de vídeo SiS5595
39	"Minitab versão 10.1" e "minitab for Windows 10.1"	Encontrar o programa de instalação da ferramenta Minitab versão 10.1 para Windows	Programa de instalação da ferramenta Minitab versão 10.1 para Windows
40	"Driver for ATI 3D Rage Pro AGP"	Encontrar o driver para Windows da placa de video ATI 3D Rage Pro AGP	Driver para Windows da placa de video ATI 3D Rage Pro AGP
41	"Driver for creative CT 4810"	Encontrar o driver para Windows da placa de som Creative CT4810	Driver para Windows da placa de som Creative CT4810

	Consulta por meio de palavras-chave	Objetivo	Consulta em língua natural
42	"Netconfig"	Encontrar o programa netconfig para Solaris.	Programa netconfig para Solaris
43	"Compactadores freeware"	Encontrar compactadores que fossem freeware	Compactadores freeware
44	"Definição /+de servidor +de curso"	Pretendi descobrir documentos na WWW que definissem "Servidor de curso"	Qual a definação de servidor de curso
45	"Definição +de servidor +de curso"		Defina servidor de curso
46	"Servidor +de curso" definição		
47	"Servidor de curso"		
48	"GIOVANNA PEREIRA GAMBARINI"	Procurei informações sobre uma nutricionista que trabalha no Núcleo de Saúde Integrada de São Carlos	Qual o e-mail de Giovanna Pereira Gambarini
49	"Glossário +de informática" Camarão	Queria saber se um glossário de informática escrito por Paulo Camarão estava disponível no formato eletrônico.	O Glossário de Informática, escrito por Paulo Camarão, está disponível gratuitamente na Internet?
50	"Paulo César Bhering Camarão"		
51	"Paulo César Bhering Camarão" "glossário +de informática"		
52	"Paulo César Bhering Camarão" glossário		
53	Camarão "glossário +de informática"		
54	"MIRAGE POP PRATA"	Descobrir o preço de uma máquina fotográfica da marca Mirage Pop Prata	Quanto custa a câmera fotográfica Mirage Pop Prata
55	"MIRAGE POP" Camera		
56	"MIRAGE POP" Câmera		
57	"Núcleo +de Saúde Integrada" "São Carlos"	Achar informações sobre o Núcleo de Saúde Integrada de São Carlos, porque não me lembrava da URL	Qual a URL do Núcleo de Saúde Integrativa de São Carlos
58	Integrada "São Carlos"		
59	Integrada "São Carlos" Núcleo		
60	Integrada "São Carlos" Núcleo Saúde		
61	Bode carneiro parapsicologia	Encontrar definições parapsicológicas de "bode" e "carneiro"	Como a parapsicologia define bode e carneiro
62	O redirecionador não conseguiu determinar o tipo de conexão.	Pretendi descobrir a causa e a solução de uma mensagem de erro que aparecia no meu computador	Por que aparece no meu computador a mensagem 'O Redirecionador não conseguiu determinar o tipo de conexão'
63	Redirecionador determinar tipo de conexão		

Discussão

O experimento realizado analisou as consultas para verificar: **1º)** se os usuários expressavam bem seus objetivos por meio de palavras-chave; **2º)** como os usuários expressariam seus objetivos em língua natural, caso esta possibilidade fosse oferecida; **3º)** se as consultas em língua natural forneciam informações que pudessem facilitar a recuperação de informação.

Antes de discutirmos os três objetivos anteriores, listamos a seguir alguns fatos inesperados que apareceram nas consultas, juntamente com possíveis explicações para eles:

1) ***Uso de palavras em inglês*** – Em alguns casos, o uso de palavras em inglês é justificável, por se tratar de algum termo que geralmente não se traduz em português (como no caso de algumas palavras técnicas). Por exemplo, a palavra *freeware* na consulta 43. Já a consulta 32 (*Beatles lyrics*) parece realmente ser uma consulta em inglês, e não em português, como havia sido solicitado;

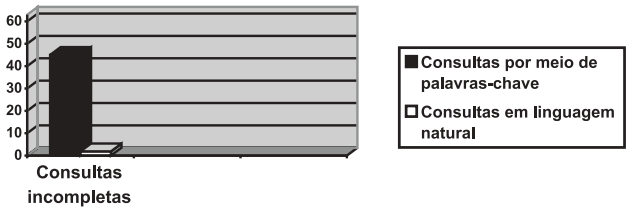
2) ***Consulta em língua natural igual ou muito semelhante ao objetivo*** – A semelhança pode tanto ser decorrente do fato de a língua natural ser uma representação do objetivo, quanto ao fato de as pessoas terem escrito a representação da língua natural depois de ter escrito o objetivo e serem influenciadas por isto;

3) ***Consulta em língua natural requerendo mais informações do que havia sido requerido nos objetivos*** – Um exemplo é a consulta de número 23, em que o usuário solicita ilustrações e críticas, e não apenas informações, como havia solicitado na consulta por palavras-chave. Uma possível causa para isto é o usuário conhecer o fato de que incluir mais palavras como palavras-chave pode gerar resultados piores do que uma consulta mais genérica;

• ***Objetivo que não retrata exatamente o que se deseja, mas sim o que o usuário tentou representar com sua consulta*** – Um exemplo é a consulta 48, sendo que, na consulta em língua natural, o usuário procura um endereço de *e-mail*, mas no objetivo ele fala de informações. Isto pode ser decorrência de o usuário saber previamente que a palavra *e-mail* algumas vezes não aparece antecedendo o *e-mail*.

Com relação ao primeiro objetivo **(1º)** – *Os usuários expressam bem seus objetivos através de palavras-chave?* –, como pode ser visto no gráfico 1, na maioria dos casos (71,43%) as consultas não trazem todas as informações declaradas como desejáveis no objetivo. Quarenta e cinco das consultas feitas com palavras-chave não continham todas as informações desejadas que haviam sido relatadas nos objetivos (por exemplo, as consultas: 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 62, 63). Ao passo que, apenas duas das representações em língua natural não continham todas as informações relatadas no objetivo como desejáveis (7 e 8).

GRÁFICO 1
Consultas incompletas



Em alguns dos casos, as consultas com palavras-chave poderiam não retornar resultados satisfatórios, por não fornecerem todas as informações desejadas, por exemplo, nas consultas 21, 22 e 23. De acordo com o objetivo, nestas consultas o usuário gostaria de encontrar informações sobre o livro ou sobre o filme "O morro dos ventos uivantes". Como isto não foi informado, o objetivo poderia ser confundido, por exemplo, com encontrar *links* para páginas que apenas vendessem o livro, e não fornecessem nenhuma informação sobre o mesmo. Neste caso em particular, o usuário especificou seu objetivo melhor em sua representação em língua natural do que no próprio objetivo. Na representação em língua natural, ele indica que quer ilustrações e críticas sobre o livro. Em outros, a falta de informações na consulta por meio de palavras-chave pode ocasionar mais do que resultados parcialmente satisfatórios, pode ocasionar resultados que sequer estão relacionados com a consulta, por exemplo, a consulta 4. Neste caso, o usuário, além de correr o risco de não encontrar o artigo, pode encontrar informações sobre outras pessoas de sobrenome "Biederman".

Para o segundo objetivo **(2º)** – *Como os usuários expressariam seus objetivos em língua natural, caso esta possibilidade fosse oferecida pelas máquinas de busca?* –, a representação* das consultas em língua natural foi feita em 71,87% dos casos no formato de perguntas (46 representações*),

* Foram 64 representações, e não 63, pois um dos usuários apresentou duas representações para a mesma consulta.

18,75% por meio de afirmações, 9,37% por meio de ordens. A tabela 2 traz os tipos de perguntas realizadas associados a exemplos e a sua freqüência.

Com relação ao terceiro objetivo do experimento **(3º)** – *As consultas em língua natural forneceram de alguma forma informações que pudessem facilitar a recuperação de informação?* –, todas as perguntas diretas deixavam claro qual era o objetivo da consulta já com a primeira palavra da frase, ou com as duas ou três primeiras (no caso das iniciadas por locuções), com exceção das iniciadas pela palavra "qual". Por exemplo, na consulta de número 13, "qual é a *homepage* da *borland* em português", a interpretação de que o resultado desejado é o endereço da *homepage* não é dada apenas pelo início da sentença. No entanto, apesar da possibilidade de interpretar de forma errada a consulta, esta consulta provavelmente poderia ser respondida corretamente por uma máquina de busca, utilizando apenas o início da sentença, já que as máquinas de busca retornam *links* como resposta, o que provavelmente não aconteceria se a aplicação fosse outra.

Já na pergunta indireta (que apareceu cinco vezes), a interpretação não é tão simples. No exemplo tratado aqui, o que se procura está no início da sentença, porém é necessário pensar melhor nas perguntas indiretas para averiguar se este é realmente um padrão. As consultas com bases em afirmações não tinham por objetivo encontrar informações sobre um tema, mas sim um objeto. Como na consulta 16, "componentes do delphi para funções matemáticas". A hipótese de que consultas deste tipo tenham sempre o objetivo de encontrar um "objeto", e não apenas informação, seria, se confirmada, um fato realmente interessante para atender bem às consultas. Bem utilizada, esta hipótese retornaria para o usuário *links* a partir dos quais pudesse fazer *download* de bibliotecas com componentes do *delphi* para funções matemáticas, e não *links* de páginas que dessem apenas informações sobre tais componentes. As consultas com base em ordens não são de fácil interpretação, é necessário conhecer o significado do verbo para interpretá-las corretamente.

TABELA 2
Interações em língua natural por meio de perguntas

Perguntas iniciadas por	Exemplos	Freqüência
Onde	Onde eu posso encontrar o programa RegClean?	13
O que é	O que é content Word?	4
Em quais	Em quais aplicações o SIMR é usado?	2
O que é e para que	O que é e para que serve reconhecimento sintático de padrão?	1
Qual	Qual o e-mail de Giovanna Pereira Gambarini?	8
Como	Como fazer análise orientada a objetos com Uml?	5
Aonde	Onde posso encontrar um ícone de página em construção?	2
Quais	Quais são as publicações de John Swales?	1
O nome do objeto procurado	O Glossário de Informática, escrito por Paulo Camarão, está disponível gratuitamente na Internet?	5
Quanto	Quanto custa a câmera fotográfica Mirage Pop Prata?	3
Por que	Por que aparece no meu computador a mensagem "O Redirecionador não conseguiu determinar o tipo de conexão"?	2

CONCLUSÕES E PESQUISAS FUTURAS

A pesquisa relatada neste trabalho tentou preencher uma lacuna na literatura relacionada a estudos que analisam consultas em português submetidas a sistemas de busca na *Web*. Em se tratando de sistemas de busca para países cujo idioma oficial é o português, só encontramos trabalhos descrevendo sistemas, analisando seu desempenho ou comparando seu desempenho com outros sistemas.

De um total de 440 *e-mails* enviados pedindo a colaboração no experimento aqui descrito, apenas 3,67% das pessoas enviaram suas consultas, e destas apenas 68,75% enviaram as consultas com objetivo e representação em língua natural. Este fato pode ser decorrente de um ou alguns dos seguintes fatores: a) o público da pesquisa foi basicamente de profissionais de

* O número de representações inclui as representações iguais para formas diferente de consultas em palavra-chave. Se a mesma representação aparece duas vezes, ela é contada duas, e não apenas uma vez.

computação, e estes profissionais em geral fazem pesquisas em inglês; b) as pessoas não têm tempo ou não se interessam em participar de pesquisas voluntariamente; c) as pessoas se esqueceram por não ser uma obrigação ou pelo período ter sido longo demais (um mês); d) as pessoas tiveram receio de relatar suas consultas pessoais.

O estudo realizado é ainda preliminar, porém aponta para os benefícios das consultas em língua natural quando contrastadas com a busca utilizando palavras-chave.

Acreditamos que a representação em língua natural é mais adequada para expressar o objetivo de um usuário do que a representação por palavras-chave. Apesar de mais detalhes causarem, em muitos casos, ambigüidade, esta ambigüidade gerada pelo excesso de informações é mais fácil de tratar do que a ambigüidade gerada por consultas simples demais. Portanto, as consultas em língua natural seriam uma forma de melhorar a precisão de máquinas de busca. No entanto, sabemos que: a) precisamos de um número maior de consultas para confirmar se a representação em língua natural realmente apresenta mais informações do que a representação em palavras-chave; b) temos de planejar melhor a coleta destas consultas para que a representação em língua natural não seja influenciada pelo objetivo; c) só poderemos saber se mais informações geram no máximo um tipo mais fácil de tratamento de ambigüidade após analisarmos exaustivamente diversos casos.

Mesmo para uma nova versão do experimento sobre as consultas em língua natural precisamos da colaboração de uma máquina de busca cedendo seus *logs*, pois precisamos fundamentar o comportamento do usuário para a primeira parte de nosso experimento, isto é, "se os usuários expressavam bem seus objetivos por meio de palavras-chave". Felizmente, conseguimos o acesso a um *log* da máquina de busca TodoBr (www.todobr.com.br), que é especializada na *Web* brasileira, e já obtivemos o acesso a um *log* da máquina de busca Tumba (www.tumba.pt) especializada na *Web* portuguesa.

Para trabalhos futuros pretendemos: 1) elaborar uma lista dos tipos de estatísticas que nos interessam e decidir entre as diversas formas de proceder este tipo de análise baseado nas análises que estudamos; 2) fazer uma análise estatística dos *logs* do TodoBr e do Tumba; 3) comparar os resultados para *Web* brasileira com os resultados para a *Web* portuguesa; 4) realizar a nova versão do experimento para verificar o que os usuários podem fornecer por meio da interação natural que pode ser explorado pelas máquinas de busca.

Além dos estudos relacionados a máquinas de busca para o português em geral, um tema interessante é discutir as características particulares dos futuros novos usuários da *Web* brasileira (pessoas comuns, que muitas vezes nunca utilizaram um computador) e que alterações estas características acarretam para as máquinas de busca e para a *Web* brasileira em geral.

AGRADECIMENTOS

Agradecemos aos que colaboraram neste estudo fornecendo suas consultas e descrevendo seus objetivos e consultas em língua natural. Agradecemos, também, a Diana Santos, por suas contribuições para a escrita deste artigo.

Artigo aceito para publicação em 23-12-2002

REFERÊNCIAS

ABDULLA, G., *et al. Characterizing WWW queries*. [S. l.] : Computer Science Department, Virginia Tech, 1997. (Technical report, 97-04).

CACHEDA, F.; VIÑA, Á. Understanding how people use search engines: a statistical analysis for e-Business. *In*: E-BUSINESS AND E-WORK CONFERENCE AND EXHIBITION, 2001. **Proceedings**... Disponível em: <http://citeseer.nj.nec.com/496769.html>. Acesso em: 20 abr. 2003.

_____. Experiences retrieving information in the world wide web. *In*: IEEE SYMPOSIUM ON COMPUTERS AND COMMUNICATIONS, 6., 2001. **Proceedings**... Disponível em: <http://citeseer.nj.nec.com/488520.html>. Acesso em: 20 abr. 2003.

JANSEN, B. J. An investigation into the use of simple queries on web IR systems. *Information Research*, v. 6, n. 1, 2000. Disponível em: <http://citeseer.nj.nec.com/420204.html>. Acesso em: 20 abr. 2003.

_____; POOCH, U. A review of web searching studies and a framework for future research. Journal of the American Society of Information Science and Technology, v. 52, n. 3, p. 235-246, 2000. Disponível em: <http://citeseer.nj.nec.com/417587.html>. Acesso em 20 de abril de 2003.

_____; SPINK, A. The excite research project: a study of searching characteristics by web users. *ASIS Bulletin*, v. 27, n. 1, p. 15-17, 2000. Disponível em: <http://citeseer.nj.nec.com/415792.html>. Acesso em: 20 abr. 2003.

_____; _____; SARACEVIC, T. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, v. 36, n. 2, p. 207-227, 2000. Disponível em: <http://jimjansen.tripod.com/academic/acad.html#ResP>. Acesso em: 20 abr. 2003.

_____. *et al*. Real life information retrieval: a study of user queries on the web. *In*: *SIGIR*, v. 32, n. 1, p. 5-17, 1998. Disponível em: <*http://*jimjansen.tripod.com/academic/acad.html#ResP>. Acesso em: 20 abr. 2003.

KIRSCH, S. Infoseek´s experiences searching the internet. *In*: SIGIR, 1998. **Proceedings**... [S. l. : s. n.], 1998.

KOBAYASHI, M.; TAKEDA, K. Information Retrieval on the web. *ACM Computing Surveys*, v. 32, n. 2, 2000. Disponível em: <http://citeseer.nj.nec.com/kobayashi00information.html>. Acesso em: 20 abr. 2003.

LAWRENCE, S.; GILES, C. L. Accessibility of information on the web. Nature, p. 107-109, 8 Jul. 1999. Disponível em: <http://www.neci.nec.com/~lawrence/papers.html>. Acesso em: 20 abr. 2003.

MARKATOS, E. P. On caching search engine: query results. *In*: INTERNATIONAL WEB CACHING AND CONTENT DELIVERY WORKSHOP, 5., 2000. **Proceedings**... Disponível em: <http://citeseer.nj.nec.com/markatos00caching.html>. Acesso em: 20 abr. 2003.

PETERS, T. The history and development of transaction log analysis. *Library Hi Tech*, v. 42, n. 11, p. 41-66, 1993.

ROSS, N.; WOLFRAM, D. End user searching on the internet: an analysis of term pair topics submitted to the excite search engine. *Journal of the American Society for Information Science*, v. 51, n. 10, p. 949-958, 2000.

SILVERSTEIN, C. Analysis of a very large AltaVista query log. [S. l. : s. n.], 1998. Disponível em: <http://citeseer.nj.nec.com/70663.html>. Acesso em: 20 abr. 2003.

SPINK, A.; XU, J. L. Selected results from a large study of web searching: the excite study. Information Research: an international electronic journal, v. 6, n. 1, 2000. Disponível em: <http://informationr.net/ir/6-1/paper90.html>. Acesso em: 20 abr. 2003.

_____ *et al.* Analysis of a very large web search engine query log. *In*: SIGIR, v. 33, n. 1, p. 6-12, 1999.

_____ *et al.* From e-sex to e-commerce: web search changes. *IEEE Computer*, v. 35, n. 3, p. 107-109, 2002. Disponível em: <http://jimjansen.tripod.com/academic/acad.html#ResP>. Acesso em: 20 abr. 2003.

_____. Searching the web: the public and their queries. *Journal of the American Society of Information Science and Technology*, v. 52, n. 3, p. 226-234, 2001. Disponível em: <http://jimjansen.tripod.com/academic/acad.html#ResP>. Acesso em: 20 abr. 2003.

WOLFRAM, D., *et al.* Vox populi: the public searching of the web. *Journal of the American Society of Information Science and Technology*, v. 52, n. 12, p. 1073-1074, 2001. Disponível em: <http://jimjansen.tripod.com/academic/acad.html#ResP>. Acesso em: 20 abr. 2003.