Saber Digital
Revista Eletrônica do CESVA
ISSN:1982-8373

GOOGLE: FERRAMENTA DE BUSCA DE INFORMAÇÃO NA WEB

PEREIRA JUNIOR, E. A.*

RESUMO

Com o crescimento da rede mundial de computadores (Internet), houve a necessidade de criação de ferramentas ou mecanismos de buscas cada vez mais inteligentes, rápidos, confiantes e precisos, dessa necessidade surgiu o sistema de busca Google, um dos mais utilizados no mundo.

Palavras-chave: Google; ferramenta de busca; Internet; informação; pesquisa.

ABSTRACT

Google, known as the most used search engine of the world; came as a necessity to fill the gap when the Computer Global Network have developed. It arose the need of tool quick and easy, but also precise and trustworthy.

Key-words: Google, search tool, Internet, information, research.

1. INTRODUÇÃO

Todos sabemos que a Internet é uma grande rede de computadores interligados, e entre seus vários objetivos, o mais importante é o acesso a informação, através da grande rede podemos obter informações sobre tudo que desejamos.

Mas a grande questão é: como obter essas informações?

A resposta é: através de mecanismos ou ferramentas de buscas na própria Internet!

Mas o que são mecanismos ou ferramentas de busca?

Ferramenta ou mecanismos de buscas são sites especializados em “varrer” todo o conteúdo da Internet e achar o que se deseja encontrar, entre essas

ferramentas poderíamos citar milhares, mas as mais utilizadas hoje são o Altavista, o Alltheweb, o Yahoo, o Msn e o Google, este último que será o foco de nosso artigo.

* Especialista em Análise e Desenvolvimento de Sistemas, Gerente de Tecnologia da Informação da FAA e leciona as disciplinas de Análise e Projeto de Sistema II, Laboratório de Programação II e Linguagem e Técnicas de Programação V no curso Superior de Tecnologia em Analise e Desenvolvimento de Sistemas de Valença - CESVA.

O Google, cuja pronúncia em português é “gugol” é a ferramenta de busca mais utilizada mundialmente, segundo estatísticas de cada 10 buscas feita na Internet, 7 são feitas através do site da Google.

2. O SEGREDO DO GOOGLE

O Google possui um sistema de atualização em sua base de dados único, esse sistema é chamado pelos desenvolvedores da Google como crawler Googlebot, é um “robô” que busca informações diariamente em tudo que é site.

Outra razão para o sucesso do Google é o sistema de PageRank (lista de popularidade do Google) que classifica os sites de acordo com a quantidade de links externos que o mesmo possui, como conseqüência, o conteúdo desse site é listado primeiro nas buscas, pois o PageRank entende que aquela página trata com maior relevância o assunto pesquisado. Além disso o Google analisa os assuntos mais pesquisados e verifica quais sites tratam aquele tema de maneira mais significativa. Para isso ele checa a quantidade de vezes que o termo pesquisado aparece na página.

Além disso o Google possui um recurso extremamente útil: o de armazenamento em cache. O Google armazena quase todas as páginas rastreadas pelo Googlebot e permite que esse conteúdo seja acessado, mesmo se essa página não existir mais. Por exemplo, suponhamos que você pesquisou um assunto e o mesmo foi encontrado em uma página, porém ao clicar no link para essa página, lhe aparece a mensagem que essa página não mais existe. Se você clicar no link “Em Cache” no resultado da busca do Google, você acessará uma cópia daquela página que está armazenada no mesmo.

Outros dois fatores que ajudaram o Google ser o mecanismo de busca mais utilizado na WWW é a simplicidade e a clareza. A combinação desses dois itens foram trabalhadas desde a sua concepção. Devido a essa filosofia é possível acessar um site de busca leve, sem poluição visual e cujas opções são facilmente localizáveis. Além de tudo já exposto, o Google ainda é capaz de realizar buscas em mais de 300 tipos de arquivos.

Uma das grandes preocupações do Google é também manter a ética em todos os países que trabalha. Por exemplo, se alguém pesquisar sobre pedofilia, encontrará textos que abordam tal assunto de maneira legal, ou seja, investigações, estudos, notícias, mas não encontrará sites com conteúdo pedófilo.

3. UM POUCO DE HISTÓRIA

A história do Google começa em 1995 com a criação de um sistema BackRub, criado na universidade de Stanford por dois estudantes de doutorado de ciência da computação: Sergey Brin, russo e Larry Page, americano. O BackRub foi sendo aperfeiçoado, e em 1998 a ferramenta ganhou o nome de Google e a empresa Google Inc. foi fundada. Quando a Google Inc. foi fundada a equipe da empresa saiu da Universidade de Stanford e foi para casa da amiga dos fundadores do Google. A medida que o Google foi crescendo, foram se juntando a equipe original nomes de peso do setor de desenvolvimento de sistemas, como engenheiros que trabalharam na Novell, Sun, Apple.

4. POR DENTRO DAS ENTRANHAS DO GOOGLE

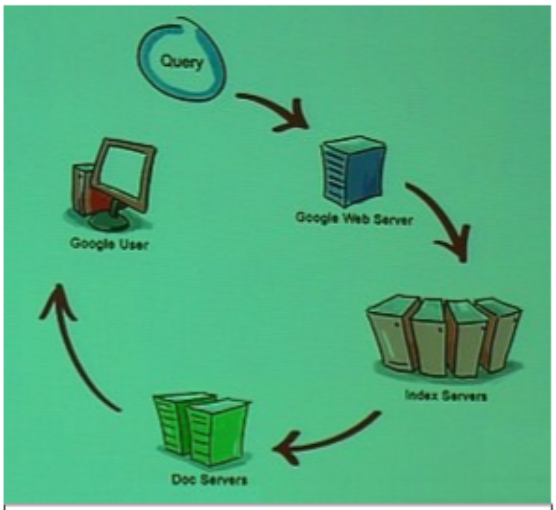
Até aqui falamos de forma superficial sobre essa maravilhosa ferramenta de busca chamada Google, agora iremos nos aprofundar mais sobre como essas pesquisas são realizadas.

Como já citado o crawler GoogleBot é responsável por varrer toda a Web e indexar seu conteúdo, para esse processo é dado o nome de *crawling*(engatinhar), em que é necessário manter o instável equilíbrio entre o tamanho do que se deseja armazenar e o tempo necessário para capturar essas informações. As necessidades

de largura de banda são tamanha, que durante esse processo *crawling* o Google pode fazer o servidor do site que está sendo vasculhado cair.

Um outro trabalho nessa fase é coibir a ação dos spammers. O conceito de spammers para o Google é um pouco diferente do que estamos acostumados.

Spammer para o Google é aquele que arruma uma forma de enganar o algoritmo de classificação do sistema, colocando um site obscuro e pouco visitado em posição de destaque PageRank da ferramenta.



Ciclo do atendimento a uma consulta no Google

A fase seguinte do Google é indexar as informações capturadas, um trabalho que dura vários dias rodando em milhares de computadores. São indexados mais de 20 bilhões de documentos, entre páginas Web, imagens, notícias e mensagens da Usenet. Nesta fase existem também as duras tarefas de eliminar duplicidade de informações e de realizar a compressão dos dados armazenados.

Durante a indexação também é calculado o PageRank de cada página, que é um número que independe das consultas (queries) feitas pelos usuários e tem a ver, isso sim, com o número de vezes que determinada página Web é linkada por outras. Uma vez que a Web já foi vasculhada e indexada, já é possível atender às consultas que nós internautas fazemos ao Google.

Para encontrar no meio do gigantesco banco de dados do Google as páginas relevantes em resposta a uma consulta, a consulta primeiro passa pelo servidor Web do Google, depois pelos servidores de índice e em seguida pelos servidores de documentos, para só então serem entregues as respostas ao usuário. A medida de relevância de cada hit (hit = cada site devolvido como resposta) é calculada com base em fatores dependentes da consulta e fatores independentes dela.

5. COMO UMA FERRAMENTA DESSA COMPLEXIDADE CONSEGUE DOMINAR O MERCADO SENDO GRATUITA?

A resposta para a pergunta é simples: ANÚNCIOS.

Também conhecidos como Google Ads, que para a empresa é o segundo sistema mais importante depois do mecanismo de busca, a implementação desse sistema é tão desafiador como a própria pesquisa por palavras-chaves, só que tem o complicador da semântica transacional que a grosso modo é uma forma de empacotar uma série de diferentes transações num conjunto de banco de dados de modo que seja vista pelo usuário como uma operação única. Para nós usuários dessa ferramenta de busca parece “mágica” receber uma lista gigantesca de sites com relevância para o que estamos buscando, e a ferramenta vai além disso, com base nessa busca o sistema exibe na tela de resposta diversos anúncios que têm a ver com o assunto pesquisado. Sendo assim uma consulta desencadeia uma série de operações nos bancos de dados que, até ser atendida, precisa ser vista pelo usuário como um bloco funcional único, ou seja, o que interessa para nós usuários é a resposta a nossa pesquisa e alguns anúncios relacionados com a mesma.

Para colocar os Googles Ads na página que o internauta está vendo é preciso que o sistema “entenda” o que está sendo tratado na tal página. Isto pressupõe que o conteúdo não está em inglês, ou seja, é necessário uma etapa de tradução. Mas essa tradução não é mostrada ao usuário, ela é feita internamente, no sistema do Google, para que seja possível que os algoritmos entendam qual o assunto que está sendo tratado na página.

Uma das saídas inteligentes para essas traduções é transformar este problema numa questão de modelagem estática e, em cima disso, partir para uma fase de “treinamento” dos algoritmos, alimentando-os com toneladas de expressões traduzidas prontas.

A partir de uma detalhada descrição probabilística do processo de tradução, utiliza-se um "corpus", ou seja, um vasto conjunto de pares de expressões, uma delas no idioma original e a outra sendo sua tradução para o inglês, que é a língua-mãe do Google. Com base nestes pares, é preciso encontrar as palavras que se relacionam à tradução e realizar o alinhamento das sentenças, criando aos poucos um modelo log-linear de tradução para o idioma em questão. Ao longo da

preparação para o evento, o Google andou também fazendo lá suas pesquisas internas e chegou à conclusão que o tamanho do corpus influi nos resultados. Quando se duplica o tamanho do corpus, os resultados melhoram cerca de 0,5%.

E não podemos esquecer que este processo precisa ser repetido para cada um dos idiomas contemplados pelo Google, quais sejam: africânder, albanês, alemão, amárico, árabe, armênio, azerbaijano, basco, bengali, bielo-russo, bihari, bósnio, bretão, búlgaro, cambodjano, catalão, cazaque, chinês (simplificado e tradicional), coreano, córsego, croata, dinamarquês, eslovaco, esloveno, espanhol, esperanto, estoniano, faroês, finlandês, francês, frísio, galego, galês, galês da Escócia, georgiano, e mais de 5 dúzias.

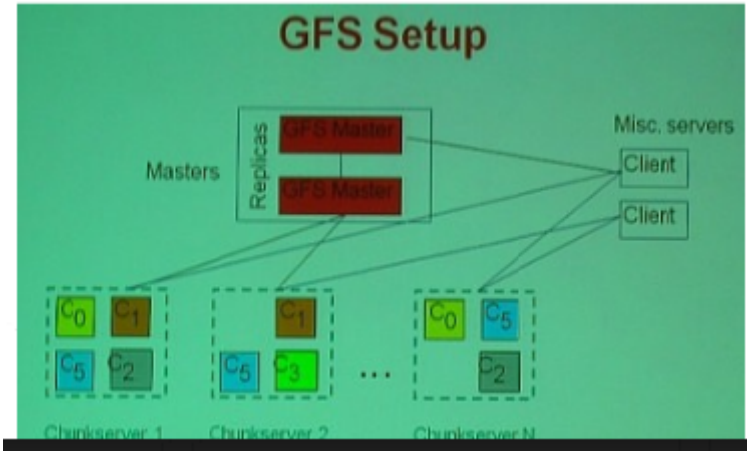
Na hora de traduzir para valer, é feita uma análise de cada frase para descobrir a probabilidade máxima de casamento com um dos pares de expressões com que foi treinado o sistema.

6. SISTEMAS DISTRIBUÍDOS NO GOOGLE

O sistema de armazenamento de informações do Google, denominado GFS (Google File System) atende a requisitos únicos e jamais antes encontrados na História da Computação:

1) Altíssima largura de banda para leitura e escrita;
2) Confiabilidade sobre uma matriz de milhares de nodos;
3) Opera em sua maioria com grandes blocos de dados;
4) Necessita de uma operação distribuída eficiente.

Cabe ressaltar que o sistema operacional utilizado pelo Google é o Linux.

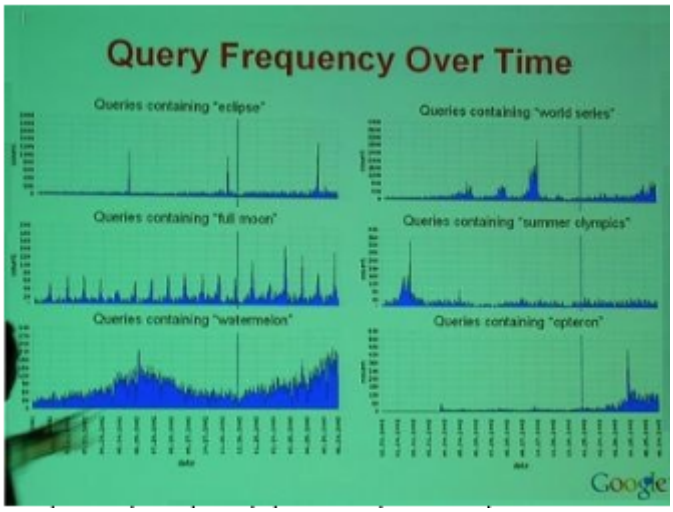


Na figura ao lado os “masters” gerenciam meta-dados, ou melhor dizendo dados a respeito de dados. A transferência de informações se dá diretamente entre os "Clients" e os "ChunkServers".

Os arquivos são quebrados em pedaços (chunks) de 64 MB cada. Com relação ao uso do GFS pelo Google, funcionam nele mais de 50 clusters, cada um com mais de 1000 computadores. São formados pools com mais de 1000 clientes cada um, num total de mais de 1 Petabyte de arquivos. A carga de leitura e escrita é de mais de 5 GB por segundo. E tudo isso na presença de frequentes falhas de hardware, ou seja, com esses dados podemos concluir que o sistema de arquivo Google é poderoso para aguentar tais circulação de dados.

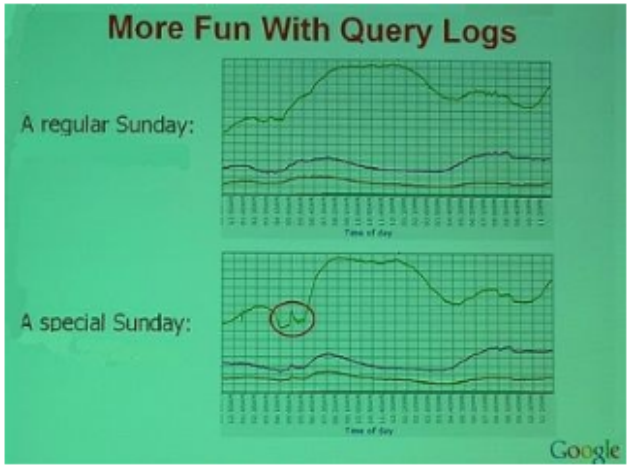
O GFS é um sistema de armazenamento distribuído de alta confiabilidade capaz de crescer até a escala de Petabytes mantendo esses dados guardados em *chunks* de 64 Mb, armazenados em discos espalhados por milhares de computadores, esses *chuncks* são armazenados em no mínimo três máquinas diferentes, o que torna o processo de perda de dados quase mínima no caso de falha de hardware.

Com toda essa tecnologia em mãos, a equipe do Google tem o poder de gerar estatísticas sobre acessos em todos os países onde o mecanismo de busca é utilizado, através da ferramenta Sawmill, é possível fazer a análise de logs e analisar os dados residentes nela e o acesso da massa de internautas.



Na figura ao lado podemos ver estatisticamente os picos de busca relacionados a assuntos por data, quando algum acontecimento importante se dá, as consultas atingem um pico maior de buscas, isso nos deixa claro que o internauta recorre ao Google para saber o que realmente está acontecendo sobre determinado assunto.

Como mostra o primeiro gráfico da figura, os picos de buscas relacionadas ao tema “eclipse” se dá justamente em época que acontecerá o eclipse.



Na figura ao lado, temos a comparação entre dois domingos, um Domingo comum, e outro um Domingo especial na Alemanha. E o que faz desse Domingo um Domingo especial? Jogo da seleção alemã na copa do mundo. Note-se no pequeno círculo vermelho a queda dos acessos ao Google durante o primeiro e o segundo tempo. No intervalo da partida, porém, que os alemães acessaram bastante, querendo saber coisas sobre o jogo, provavelmente.

7. NECESSIDADES COMPUTACIONAIS DO GOOGLE

Suas necessidades computacionais se baseiam no tripé:

- Mais dados
- Mais consultas
- Melhores resultados

É fácil explicar esse “tripé”:

Mais dados por que a Web não pára de crescer,

Mais consultas por que o número de usuários não param de aumentar, e quando mais os usuários estiverem satisfeitos com o serviço, a tendência é que o sistema cresça cada vez mais e mais rápido;

Melhores resultados por que a equipe do Google permanece continuamente pesquisando novas maneiras de atender as consultas da forma ainda mais rápida e precisa.

8. HARDWARE

Em relação ao *hardware* utilizado pela equipe da Google, a preferência é para servidores de baixo custo, o desempenho de uma máquina “hiper-poderosa” não interessa a eles. O grande problema Google é quebrado em pequenos problemas e particionados em vários *Threads* de processamento, o que está de pleno acordo com a tendência moderna de chips multicore (múltiplos núcleos).

Um outro aspecto para a não utilização de “hiper-máquinas” é que torna os programadores preguiçosos, e é levado em consideração que até mesmo as plataformas mais confiáveis não estão imunes a falhas, tornando assim clara a necessidade de um software tolerante a falhas que pode funcionar em componentes de baixo custo.

9. DICAS PARA USO DO BUSCADOR GOOGLE

Quem sabe aproveitar todas as opções do Google certamente encontrará o que deseja, por mais complexo ou por mais desconhecido que o assunto seja. Se os recursos certos forem usados, as chances de encontrar algo que seja interessante à pesquisa aumentam consideravelmente.

Cálculos

Fazer cálculos no Google é simples. Digite, por exemplo, 2 + 2, 18 * 3, 14 / 8 ou 4 − 3 e veja o que acontece. O Google consegue realizar desde operações básicas até as mais complexas. Basta digitar o tipo de cálculo desejado. Veja a lista:

- 5 ^ 3 —> faz 5 elevado a 3 (53)
- sin(45 degrees) —> faz o seno de 45 ł
- tan(45 degrees) —> faz a tangente de 45 ł
- cos(45 degrees) —> faz o coseno de 45
- sqrt (90) —> faz a raiz quadrada de 90
- ln (13) —> faz o logaritmo base e de 13
- log (1,000) —> faz o logaritmo base 10
- 50! —> faz o fatorial de 50
- 4th root of 64 —> faz o cálculo da quarta raiz de 64 łł

O degrees não é obrigatório. Digite-o somente quando desejar o valor em graus. Sem o degrees, o valor é fornecido em radianos.

Para 1, deve-se usar st em vez de th. O mesmo vale para 2, onde deve-se nd e 3, onde deve-se usar rd. Para 4 e os demais números, deve-se usar th.

Você não precisa usar cada operação por vez. É possível criar uma equação. Por exemplo, (14 + 554) * ln(13) / tan(90) + 1. O Google dará como resultado −729.197942.

Conversões

É possível fazer conversões no Google. Veja a lista de conversões:

- 50 miles in km
- Faz 50 milhas em quilômetros
- 10 kg in lb
- Faz 10 quilos em libras
- 30 cm in ft
- Faz 30 centímetros em pés
- VI in arabic numerals
- Transforma VI em número arábico (o que utilizamos hoje em dia)
- 2004 in roman numerals
- Transforma 2004 em números romanos
- 9 hours in minutes
- Transforma 9 horas em minutos
- 365 days in hours
- Transforma 365 dias em horas

Em todos os casos, é possível que você faça as operações de modo contrário. E há outras conversões. Basta saber os nomes das medidas em inglês e experimentar no Google. No lugar dos valores, você pode usar equações. Por exemplo, 10/5+459 in roman numerals. O Google mostrará CDLXI. Outra maneira de fazer este tipo de conversão é escrevendo em formato de pergunta a medida desejada, em inglês, como nos exemplos abaixo:

How many cm are in 40 km? Quantos centímetros há em 40 km? How many miles are in 9041 cm?
Quantas milhas há em 9041 centímetros? How many hours are in 8 days?
Quantas horas há em 8 dias?

Operadores Avançados

Algumas dicas que servem para a maioria dos buscadores, incluindo o Google, é usar apenas palavras chaves na sua busca, ao invés de buscar, por exemplo, Golpe do Estado busque por Golpe Estado. O Google e alguns poucos outros tem uma melhor vantagem, ele faz um filtro de busca, retirando informações pequenas, como de, da/do, com, dessa forma não é totalmente necessário fazer o filtro manualmente, porém a pesquisa pode se tornar um pouco mais rápida.

Função	Exemplo	Descrição
Pesquisa Exata	"Google Search"	Procura pela ocorrência EXATA (com as palavras agrupadas) de Google Search.
Filtrar Resultado	Google -Search	Filtra o resultado removendo todos os que possuem Search como resultado.
Busca Alternativa	Google (Search OR Groups)	Ao invés de OR.
Curingas	"Google * tem ótimas opções"	Troca o asterisco por uma palavra ou frase desconhecida.
Procurar num Site	Google site:pt.wikipedia.org	Procura pela palavra Google no site pt.wikipedia.org.
Buscar por tipo de arquivo	Google filetype: PDF	Procura a palavra Google em arquivos com extensão PDF.
Combinar Informações	filetype: PDF site:pt.wikipedia.org	Procura por arquivos de extensão PDF no site da pt.wikipedia.org.
Buscando pelo URL	inurl:wikipedia	Procura wikipedia no URL do site.
Buscando pelo Texto	intext: wikipedia	Procura pelo texto wikipedia no conteúdo do site, você pode simplificar este uso digitando somente wikipedia.
Buscando Conceitos	define: wikipépia	Define a palavra Wikipédia.
Palavras Chaves	keyword: wikipedia	Procura na METATAG do site por Wikipédia isto algumas vezes podem ser mais funcional.
Cache	Cache:www.google.com	Vê a página www.google.com em cache.
Titulo	intitle:google wikipedia	Procura páginas que tenham Google e/ou wikipedia no título da página.

10.CURIOSIDADES SOBRE O GOOGLE

• A empresa Nielsen//NetRatings que mede a audiência de sites em 11 países divulgou no ano de 2005 uma relação de usuários de sistemas de buscas. O Google se encontra em primeiro lugar com 153 milhões no mês de agosto, contra 150,6 milhões do MSN e 146,5 milhões do Yahoo!.

• Atualmente o Google é o terceiro site mais acessado do mundo e o quarto no Brasil, ficando somente atrás dos portais UOL, IG e Terra.

• A 11ª edição do dicionário em inglês Merriam-Webster's Collegiate Dictionary passou a classificar a palavra "Google" como um verbo transitivo cujo passado é "googled"[2]. Segundo a definição, Google significa "usar a ferramenta de buscas Google para obter informações na world wide web".

• Nos EUA o Google é responsável por 64% das buscas na Internet.

• A expressão googol surgiu de um fato um tanto curioso, o matemático Edward Kasner questionou o seu sobrinho de 8 anos sobre a forma como ele descreveria um número grande - um número realmente grande: o maior número que ele imaginasse. O pequeno Milton Sirotta emitiu um som de resposta que Kasner traduziu por "googol".

CONCLUSÕES

Podemos concluir que a ferramenta de busca Google domina a Internet pelo fato de ser uma ferramenta simples, porém robusta, se destacando das demais pela rápida resposta a requisição de buscas, e também possuindo uma interface amigável, sem “poluições”, o que torna a navegação por essa ferramenta muito mais agradável.

Outro ponto interessante que cabe ressaltar é a forma como seus idealizadores encontram soluções simples para problemas complexas, indo as vezes, no inverso do que se estuda na academia, exemplo para essa afirmação é a utilização de computadores normais para gerir toda a informação que essa ferramenta produz.

A Google Inc. possui dezenas de outros serviços, citando os mais importantes temos:

AdSense: ajuda você a gerar receita para seu site, com anúncios relevantes, ou seja, eles aparecem de acordo com o conteúdo que seu site oferece.

Analytics: é um serviço do Google que mostra a percentagem de visitas por mês, o navegador dos visitantes, sistema operacional, a localização geográfica, etc.
Blog Search: é um serviço do Google especializado para buscas em blogs. Os bots do Blog Search parecem ser mais rápidos do que o Googlebot padrão, pois

atualizações feitas em blogs, muitas vezes se tornam disponíveis em poucas horas ao contrário das semanas levadas pelo Googlebot padrão.

Gmail: Servidor de e mail gratuito que atualmente oferece ao usuário do seu serviço o espaço de 2 Gb para recebimento de mensagens, arquivos entre outros.

YouTube: é um sistema que permite o envio de videos para a internet armazendo-o, serviço que não foi criado pelo Google mas foi adquirido pelo mesmo por um valor de U$1,6 bilhão. O Youtube é líder na área de vídeos on-line.

Google Earth: Emite imagens satélites de várias cidades do planeta, incluindo estradas, estações de metrô, etc.

REFERÊNCIAS BIBLIOGRÁFICAS

DANTAS, M. **Dominando o Google**. Rio de Janeiro: Editora BrasPort, 2005.

MONTEIRO, R.V. **Google Adwords: a arte da guerra**. Rio de Janeiro: Editora BrasPort, 2007.

VISE, D.A.; MALSEED, M. **Google: a história de mídia e tecnologia de maior sucesso em todos os tempos**. São Paulo: Editora Rocco, 2007.

PORTAL Google. Disponível em: <http://www.google.com/corporate/history.html>. Acesso em 06 set. 2007.

BLOG Oficial do Google no Brasil. Disponível em: <http://googlebrasilblog.blogspot. com/>. Acesso em 06 de set. 2007.

WIKIPÉDIA. Disponível em: <http://pt.wikipedia.org>. Acesso 06 de set. 2007.