Universidade do Porto
FEUP Faculdade de Engenharia

Information Retrieval Techniques in Commercial Systems

Disciplina de **ARMAZENAMENTO E RECUPERAÇÃO DE INFORMAÇÃO**

Professor Mark Sanderson

Cristina Henriques

Jorge Meneses Freitas

PORTO
10 de Junho de 2001

1. Introdução

a. Objectivo

O objectivo deste estudo é perceber como funcionam e caracterizar três sistemas de motores de busca. Praticamente, desde o inicio da web, que estes programas apareceram e tiveram uma adesão elevada. Mas a pergunta põe-se: Será que são que os motores de busca são eficazes? Será que encontram tudo aquilo que deviam encontrar, ou apenas parte?

Também é sabido, que a maior parte das pessoas apenas lê os primeiros resultados da pesquisa, e assim, a ordenação dos resultados torna-se extremamente importante. Como são apresentados os resultados? Aparecem primeiro os documentos mais relevantes para o utilizador ou existem formas comerciais de colocar determinadas referências em primeiro lugar?

São estas e outras perguntas que se irá tentar responder com este estudo.

Este estudo baseia-se unicamente na leitura dos prospectos comerciais e na caracterização dos sistemas a partir deles.

b. Sistemas Escolhidos

Os sistemas escolhidos são:

- Altavista: www.altavista.com
- Google: www.google.com
- WebTop: www.webtop.com

A escolha foi feita nestes motores, essencialmente por duas razões:

- Popularidade na utilização, uma vez que são amplamente conhecidos por quem utiliza a web (essencialmente o Altavista e o Google)
- Descrição técnica do sistema, pois encontrou-se facilmente documentação nos sites respectivos sobre as formas de pesquisa utilizadas.

2. Descrição dos Sistemas

a. Altavista

A empresa

A empresa Altavista foi fundada em 1995. Os seus fundadores, referem-se a esta palavra com sendo algo que representa "uma vista de cima". Esta empresa surgiu a partir dos resultados conseguidos por uma equipa de cientistas da Digital, no que toca à indexação de palavras que estavam nas páginas da web. Está sedeada nos Estados Unidos da América, no estado da Califórnia na cidade de Palo Alto.
Actualmente, tem cerca de 550 funcionários e vende o seu motor de pesquisa a mais de 1000 empresas orientadas para a Internet. O controle da empresa é feito através de um grupo empresarial chamado CMGI.

O método de Pesquisa

Basicamente, funciona numa base de indexação. Existem programas que são denominados como robots, crawlers ou spiders que estão continuamente a navegar na web, começando por uma página, fazendo a leitura da mesma e seguindo para os links que lá estão apontados.
Ao fazer essa passagem, vão registando texto numa base de dados de elevada dimensão, onde existe o endereço da página e um texto descritivo da mesma. Quando alguém faz uma pesquisa, colocando lá palavras, o que é feito é uma procura na base de dados das palavras que lá estão contidas, apresentando os links para as respectivas páginas.
Existem algumas características chave deste sistema:

- Tem um dicionário de 500.000 entradas que serve para verificar se um conjunto de palavras forma uma frase. Ou seja, quando se escreve o nome de uma pessoa, género Jorge Meneses Freitas, o sistema percebe que não deve tratar a pesquisa como três palavras individuais, mas sim como uma frase de três palavras.
- Faz a distinção entre resultados encontrados na mesma página, ou seja, só retorna uma vez o documento, mesmo que exista várias referências ao texto pesquisado naquele documento.

- Projecta aquilo que será mais útil para o utilizador, fazendo uma procura na base de dados construída, e ordenando os resultados através de alguns critérios mais finos:
    - Começa por verificar a informação relativa ao título da página; se existir a comparação correcta, é logo apresentada como informação significativa.
    - Se encontrar o texto procurado no início do documento, é dada relevância a essa página.
    - Se encontrar o texto em páginas que são ricas em texto significativo, é-lhes dada também relevância.
    - As páginas que tenham muitos links para outras páginas com conteúdos semelhantes, também são postas em evidência.

O índex que existe actualmente é enorme, contendo chaves para centenas de milhões de páginas.

Existe o conceito de "Listagem Paga", em inglês "Sponsored Listings". Isto funciona quando se pesquisa artigos ou serviços comerciais, e as empresas pagam para que a sua referência surja na primeira página. É de referir que isto só acontece quando o resultado da pesquisa é exacto relativamente ao texto que foi pago pela empresa.

Aquilo que aparece debaixo do título do link, é a descrição. Por defeito, consiste nos primeiros 180 caracteres que aparecem na página. Mas se houver meta - informação, através de "description metatag", o robot irá registar o texto aí registado.

O Altavista tem uma ferramenta de tradução, chamada "BabelFish". O que permite é traduzir uma página de Internet de uma linguagem que o sistema conheça para outra que ele também conheça. As linguagens conhecidas são Inglês, Francês, Alemão, Espanhol, Português, Russo, Chinês, Coreano e Japonês).
O resultado da tradução de Português para Inglês, é razoável, falhando apenas algumas palavras. De Inglês para Português, é também razoável, mas percebe-se que não é Português correcto, ou seja, é Português "Brasileiro".

b. Google

Como surgiu?

Dois estudantes avançados da Universidade de Stanford (EUA), Sergey Brin e Lawrence Page, desenvolveram um projecto de um motor de busca. Em 1998, fundaram a empresa Google.

O método de pesquisa

Funciona com uma lógica de indexação semelhante à do Altavista. Existem robots, denominados crawlers, que estão continuamente a navegar na web a registar os sites que existem e o seu conteúdo. A diferença mais significativa é o método para a ordenação dos resultados que aparecem.

O método que é utilizado chama-se "PageRank". Basicamente, consiste na lógica das citações, ou melhor, "links" das outras páginas para esta. O que o sistema faz, é calcular um valor que a página tem com base em quantos links são feitos para aquela página. Também existe a lógica de ter bastante valor de ranking quando uma página com o PageRank muito elevado faz um link para aquela página. Se imaginarmos uma página como o Yahoo! a referir directamente uma determinada página, é porque deve ser credível e como tal deve ser também muito valorizada. Ou seja, é a lógica de ser referido por muitos web sites ou por web sites altamente conhecidos.
Fazendo a citação directa do documento de suporte [1], PageRank é definida da seguinte forma:

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

Outra técnica usada paralelamente à da PageRank é a da AnchorText. A lógica é a de registar não só o link e dar-lhe pontos a esse destino, mas também de registar o texto que esse link tem. A vantagem apresentada é que esse texto, muitas vezes, é uma melhor descrição para a página em causa.
Outra característica deste sistema é que dá mais "pontos" às palavras que estão maiores do que o normal ou em bold.

c. WebTop

Como surgiu?

A Webtop é um site que foi lançado pela Dialog Corporation e esta ainda o detêm. O arranque do sistema deu-se em Dezembro de 1999. A sua forma base de pesquisa, o Linguistic Inference, foi desenvolvido na Universidade de Cambridge, pelo Prof. Martin Porter, pelo Prof. Keith Van Rijsbergen e pelo Prof. Stephen Robertson.

Porque surgiu?

A Dialog é possuidora, desde há muitos anos para cá, de bases de dados de muitos temas de informação que estão devidamente estruturados e que o seu acesso é vendido. Através de ferramentas inteligentes de negócio, as pessoas que têm acesso podem pesquisar essa informação nessas estruturas de informação.
Nos dias que correm, o mercado solicita que a pesquisa de informação esteja disponível a todos os elementos da organização, mesmos que estes sejam aprendizes nessas mesmas pesquisas. Assim, surge a necessidade de uma ferramenta, facilmente distribuível por muitos utilizadores dispersos geograficamente. Assenta numa plataforma Web e de livre acesso, é criado o WebTop.

O método de pesquisa

O método de pesquisa de base é o Linguistic Inference. Este parte do princípio que o utilizador a fazer uma pesquisa não sabe exactamente como há-de procurar o que precisa e que começam a procura com uma quantidade elevada de ambiguidade. Esta tecnologia, está assente em cinco directrizes:

- Data Collection: Continuamente é feita a revisão do conjunto de informação para identificar nova e informação actualizada. É um processo que permite obter informação de formatos pré-definidos e automaticamente seguir os links lá apresentados (semelhante ao Altavista e ao Google).
- Concept Extraction: Identifica e extrai conceitos a partir do conjunto de informação. Analisa a informação que está a entrar, identifica importantes palavras e frases, identifica chaves de relação entre palavras e frases e guarda toda a informação no Concept Map (CMAP), que fisicamente é um ficheiro de indexação invertido. Existem duas formas de fazer esta indexação:
    - High Recall Extraction: Todas as palavras são indexadas, ou seja, fazendo uma pesquisa por "desporto futebol", todos os documentos que tenham uma das duas palavras são retornados.
    - High Precision Strategy: Indexa apenas todos os conceitos importantes (definidos de forma estatística), para que quando se pesquisa "desporto futebol", só sejam retornados os documentos que tenham estas palavras como tema central. É esta forma que o WebTop utiliza.
- Interest Recognition: Identifica e conceptualiza a essência da necessidade inicial de informação de um utilizador, através de dedução de um ou mais conceitos a partir do interesse do mesmo. Esta informação é guardada em algo que se chama Personal Concept Profile, PCP.
- Probabilistic Concept Correlation: Correlaciona conceitos definidos pelo utilizador ou pelas suas acções com os conceitos extraídos dos documentos e das bases de dados. Em termos técnicos, quer dizer que correlaciona o PCP com o CMAP. Neste contexto, existem dois paradigmas da recuperação da informação:
    - Frequência Estatística de Palavras, ou seja, como e qual a frequência das palavras que aparecem no documento.

  - Probabilistic Retrieval Paradigm, ou seja, atribui pesos a cada palavra dinamicamente para identificar a chave dos conceitos de relevância para o utilizador. É esta que é usada pelo WebTop.
- Interest Refinement: Interage com o utilizador para o ajudar a redefinir os seus interesses através de sugestão de outras palavras, frases e documentos. O valor da linguagem natural é também realizado neste processo. Através da disponibilidade de mecanismos para o utilizador redefinir, educar ou simplesmente recordar as suas noções de relevância, o utilizador pode mover-se de um perfil para outro.

3. Conclusões

Perguntar se uma forma de pesquisa é melhor do que outra é o mesmo que perguntar se o futebol é melhor que o basquetebol. É uma questão de gosto. Se preferimos uma forma de pesquisa centrada no utilizador, devemos apontar para a Linguistic Inference e daí leva-nos ao WebTop.

Se preferirmos, uma lógica orientada unicamente à indexação, vamos cair no Altavista e no Google.

As grandes vantagens do Google sobre o Altavista são:

- É mais rápido, tanto no carregamento inicial como na apresentação de resultados
- Não tem publicidade
- O método de pesquisa revela-se mais eficaz, pois a lógica do PageRank funciona bastante bem.

A grande vantagem do Altavista sobre o Google, é a facilidade de tradução, em n línguas, de qualquer página que esteja na web.

Recomendamos que, de uma forma geral, seja utilizado o Google. Apenas em casos específicos, como a necessidade de tradução de páginas, então deve-se usar o Altavista.

4. Referências Bibliográficas

- Altavista - *Frequent Questions from Searchers,* http://help.altavista.com/search/faq

- Altavista - *Search Help*, http://www.altavista.com/sites/help/search/search_help

- Google - *Porquê usar o Google?,* http://www.google.com/intl/pt/why_use.html

- Sergey Brin e Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

- Best of the Web 1994 – Navigators, http://botw.org/1994/awards/navigators.html

- Bill Clinton Joke of the Day: April 14, 1997, http://www.io.com/~cjburke/clinton/970414.html.

- Bzip2 Homepage, http://www.muraroa.demon.co.uk/

- Google Search Engine, http://google.stanford.edu/

- Harvest, http://harvest.transarc.com/

- Mauldin, Michael L. Lycos Design Choices in an Internet Search Service, IEEE Expert Interview http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm

- The Effect of Cellular Phone Use Upon Driver Attention, http://www.webfirst.com/aaa/text/cell/cell0toc.htm

- Search Engine Watch, http://www.searchenginewatch.com/

- RFC 1950 (zlib), ftp://ftp.uu.net/graphics/png/documents/zlib/zdoc-index.html

- Robots Exclusion Protocol, http://info.webcrawler.com/mak/projects/robots/exclusion.htm

- Web Growth Summary, http://www.mit.edu/people/mkgray/net/web-growth-summary.html

- Yahoo!, http://www.yahoo.com/

- Serge Abiteboul and Victor Vianu, *Queries and Computation on the Web*. Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.

- Ben H. Bagdikian. *The Media Monopoly*. 5th Edition. Publisher: Beacon, ISBN: 0807061557

- Junghoo Cho, Hector Garcia-Molina, Lawrence Page. *Efficient Crawling Through URL Ordering.* Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.

- Luis Gravano, Hector Garcia-Molina, and A. Tomasic. *The Effectiveness of GlOSS for the Text-Database Discovery Problem.* Proc. of the 1994 ACM SIGMOD International Conference On Management Of Data, 1994.

- Jon Kleinberg, *Authoritative Sources in a Hyperlinked Environment*, Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.

- Massimo Marchiori. *The Quest for Correct Information on the Web: Hyper Search Engines.* The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.

- Oliver A. McBryan. GENVL and *WWWW: Tools for Taming the Web. First International Conference on the World Wide Web.* CERN, Geneva (Switzerland), May 25-26-27 1994, http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps

- Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Manuscript in progress. http://google.stanford.edu/~backrub/pageranksub.ps

- Brian Pinkerton, *Finding What People Want: Experiences with the WebCrawler.* The Second International WWW Conference Chicago, USA, October 17-20, 1994. http://info.webcrawler.com/bp/WWW94.html

- Ellen Spertus. *ParaSite: Mining Structural Information on the Web.* The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.

- *Proceedings of the fifth Text REtrieval Conference (TREC-5).* Gaithersburg, Maryland, November 20-22, 1996. Publisher: Department of Commerce, National Institute of Standards and Technology. Editors: D. K. Harman and E. M. Voorhees. Full text at: http://trec.nist.gov/

- Ian H Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* New York: Van Nostrand Reinhold, 1994.

- Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. *HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering.* Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.