Bruno Viana Rezende

Metodologia para otimização de sítios para motores de busca

Belo Horizonte

2004

Bruno Viana Rezende

Metodologia para otimização de sítios para motores de busca

Dissertação apresentada ao programa de pós-graduação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais (ECI-UFMG).
Orientador: Marcello Peixoto Bax

Belo Horizonte

2004

Rezende, Bruno Viana.
R467m Metodologia para otimização de sítios para motores de busca [manuscrito] / Bruno Viana Rezende. – 2004.
64 f. : il.

Orientador: Marcello Peixoto Bax.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências bibliográficas: f. 62-64

1. Ciência da informação - Teses 2. Sistemas de recuperação da informação – Teses 3. Internet – Teses I. Título II. Bax, Marcello Peixoto III. Universidade Federal de Minas Gerais. Escola de Ciência da Informação.
CDU: 004.738.5

Catalográfica: Biblioteca Etelvina Lima, Escola de Ciência da Informação da UFMG

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"METODOLOGIA PARA OTIMIZAÇÃO DE SÍTIOS EM MOTORES DE BUSCA".

Bruno Viana Rezende

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Mestre em Ciência da Informação"**, linha de pesquisa **"Gestão da Informação e do Conhecimento (GIC)"**.

Dissertação aprovada em: 15 de dezembro de 2004.

Por:

Prof. Dr. Marcello Peixoto Bax –ECI/UFMG (Orientador)

Prof. Dr. Nívio Ziviani –DCC/UFMG

Profa. Dra. Lídia Alvarenga –ECI/UFMG

Profa. Dra. Gercina Ângela Borém de Oliveira Lima –ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Maria Eugênia Albino Andrade
Coordenadora

Versão final Aprovada por

Prof. Marcello Peixoto Bax
Orientador

Agradecimentos

Ao meu pai e à minha mãe pelo apoio em todos esse anos de convivência.

Aos meus irmãos, Sérgio, Silvia e Letícia, pela paciência e amizade.

Aos demais familiares pelos momentos agradáveis que passamos juntos.

Aos amigos, sempre dispostos a tomar uma cerveja e jogar conversa fora.

Ao meu orientador pelos conhecimentos e auxílio na construção desta dissertação.

À ECI e à UFMG por me abrigar e permitir o aumento de meus conhecimentos.

Assegura-se que o mundo, abreviando as distâncias, transmitindo o pensamento pelos ares, irá unir-se cada vez mais, que a fraternidade reinará. Ai! Não acreditem nessa união dos homens. Concebendo a liberdade como o aumento das necessidades e sua pronta satisfação, alteram-lhes a natureza, porque fazem nascer neles uma multidão de desejos insensatos, de hábitos e imaginações absurdos. Não vivem senão para invejar-se mutuamente, para a sensualidade e a ostentação. Dar jantares, viajar, possuir carruagens, cargos, lacaios, passa tudo como uma necessidade à qual se sacrifica até sua vida, sua honra e o amor à humanidade, irão até matar-se na impossibilidade de satisfazê-la.

Trechos das conversas e da doutrina do "stárets" Zósima, *Os irmãos Karamazov* (1879), Fiódor Dostoievski

Sumário

Lista de tabelas

Lista de Abreviaturas

CDU – Classificação Decimal Universal

HTML – *HyperText Markup Language*

IDF – *Inverse Document Frequency*

RI – Recuperação de Informação

SEO – *Search Engine Optimization*

SRI – Sistema de Recuperação de Informação

TF – *Term Frequency*

URL – *Uniform Resource Locator*

Resumo

Apresenta uma metodologia para a otimização de sítios para motores de busca baseada em conceitos de indexação manual de documentos, assunto próprio da Ciência da Informação. Descreve o que são Sistemas de Recuperação de Informação (SRI), explica como um motor de busca (uma classe de SRI) funciona e o que é otimização de Sítios, diferenciando-a da prática de *Spam*. Como principais conclusões e contribuições deste trabalho pode-se enumerar: (1) Um sítio otimizado pela metodologia desenvolvida terá maior visitação gerada por motores de busca do que um que não tenha sido otimizado, mostrando que a otimização de sítios é uma tarefa eficaz; (2) A metodologia divide o processo de otimização entre profissionais de campos de atuação distintos, especializados em áreas específicas, ajudando a viabilizar e estruturar um serviço inovador no mercado da informação; (3) Finalmente, a dissertação ajuda a preencher a lacuna de trabalhos acadêmicos a respeito do tema e abre caminho para pesquisas futuras mais profundas.

Abstract

Presents a Search Engine Optimization (SEO) methodology based on traditional concepts of document indexing from Information Science. Describes what Information Retrieval Systems (IRS) are, how a Search Engine (an IRS class) works and what is SEO and its differences from Spam practices. It can be enumerated as the main conclusions and contributions of this work: (1) A site optimized by using the methodology developed here will have more visits generated by Search Engines, showing that Search Engine Optimization is an effective task. (2) The methodology allows the division of the process between professionals of different fields, specialized in specific tasks of SEO, what can make viable the creation of an innovative service in the information market. (3) At last, the research helps to fill the gap of academic works concerning the subject and opens opportunities for deeper future works.

1. Introdução

A recuperação de informação é uma atividade realizada originalmente em unidades de informação, tais como centros de documentação e bibliotecas, por meio da classificação e indexação realizadas previamente sobre o acervo. Com o advento da *web*, os Sistemas de Recuperação de Informação (SRI) começaram a ser utilizados para uma nova função: encontrar recursos que atendam à necessidade de informação de usuários na internet. O uso dos SRI na *web* começou cedo; exemplos de tais sistemas são o Yahoo![1] e o Altavista[2], o primeiro iniciou suas atividades em 1994, o segundo em 1995.

O objetivo dos SRI é facilitar para usuários o acesso à informação relevante, que atenda suas necessidades, com rapidez e precisão. Os SRI existentes empregam diferentes formas para tratar as informações que são apresentadas aos usuários como resposta a consultas. Uma diferença comum diz respeito a como a maior parte do processo[3] é efetuada: manualmente (por pessoas) ou automaticamente (por máquinas, softwares, etc). Esta diferença de procedimento acarreta diferenças na qualidade da recuperação e no volume de informações tratadas. Espera-se, em geral, que o tratamento manual leva a descrições de maior qualidade e, conseqüentemente, a recuperação de informação mais pertinente do que aquele feito automaticamente. O volume de informações tratadas, porém, é muito menor.

---

[1] http://www.yahoo.com

[2] http://www.altavista.com

[3] As diferentes etapas realizadas por um SRI serão apresentadas no Capítulo 2.

Na *web*, o volume de informações disponível é gigantesco, além de mudar constantemente (o Google[4] contabilizava, em fevereiro de 2003, três bilhões de páginas indexadas, em novembro de 2004 mais de oito bilhões de páginas). Esse fato torna extremamente difícil, senão impossível, a análise manual de todos os documentos disponíveis, acentuando as vantagens da abordagem automática[5]. Tal abordagem é utilizada, por exemplo, em motores de busca (em inglês *Search Engine*).

Quando procura por algo utilizando um motor de busca, o usuário expressa suas necessidades de informação em um conjunto de palavras-chave, que espera representar tais necessidades. O motor realiza a pesquisa nas páginas conhecidas, já indexadas, e apresenta como resultado referências para aquelas que têm alguma relação com o tema procurado e expresso pelas palavras-chave utilizadas. Esse resultado, geralmente apresentado por ordem de relevância, constitui-se de uma lista de URL's (*Uniform Resource Locator*), cujo número pode variar de poucas unidades até milhões de unidades. Na maioria dos casos, entretanto, apenas os 10 resultados mais relevantes são exibidos na primeira página de resultado. Assim, um sítio *Web*, almejando ser visitado por usuários interessados, deve procurar ser listado nas primeiras posições para consultas relacionadas ao seu conteúdo. Com efeito, pesquisas mostram que a maioria dos usuários ou encontra um apontador satisfatório para a informação desejada na

[4] http://www.google.com/

[5] Existem na *web*, entretanto, SRIs que se situam em um meio termo entre a abordagem manual e a automática: os diretórios. Nessas ferramentas, as informações são tratadas por seres humanos, utilizando-se máquinas apenas como auxílio, porém esses sistemas não serão considerados neste documento.

primeira página - no máximo na segunda -, ou realiza uma nova consulta, não visitando as páginas seguintes (SILVERSTEIN, 1998). O fato de um sítio, estar listado na décima primeira posição, por exemplo, pode significar deixar de receber a visita de muitos usuários (em 85% das consultas no Altavista em 1998, apenas os 10 primeiros resultados foram vistos).

Objetivando inferir a relevância das páginas em um sítio *Web* para uma dada consulta, um motor de busca utiliza evidências internas e externas. Define-se evidência interna como qualquer evidência que possa ser extraída de dentro da própria página, desconsiderando todas as outras páginas da coleção indexada. Uma evidência externa é definida como uma evidência que deve ser extraída considerando as demais páginas da coleção indexada. Por exemplo, o número de vezes que uma palavra ocorre dentro de uma página é uma evidência interna, enquanto que o número de páginas da coleção em que esta mesma palavra ocorre é uma evidência externa. Nota-se, portanto, que se as evidências internas utilizadas por um motor são conhecidas, uma página pode ser construída ou alterada seguindo um processo que vise aumentar a probabilidade desta ser considerada relevante para consultas de interesse. Mais do que isso, se evidências externas também são conhecidas, tal processo pode ser ainda melhorado. Neste documento, uma página onde o maior número possível de evidências foram adequadas conforme o processo citado é dita estar "otimizada". O processo de adequar as evidências das páginas que constituem um sítio é chamado de **otimização de sítios para motores de busca** (em inglês. *Search*

Engine Optimization – ou *SEO*). No presente trabalho de pesquisa, o processo citado será designado **otimização de sítios**.

Assim, o objetivo deste trabalho de pesquisa é a construção de uma metodologia para a otimização de sítios para motores de busca baseada em conceitos tradicionais de indexação manual de documentos, assunto próprio da Ciência da Informação. Para que tal metodologia seja legítima, é fundamental que se mantenha dentro da ética da *web*, não utilizando técnicas de *spam*, o que será explicado mais adiante no texto.

A metodologia construída demonstra que a otimização de sítios é uma tarefa eficaz e viável, que agrega valor aos sítios, aumentando o número de visitas geradas por motores de buscas. A metodologia prevê a divisão do processo de otimização entre profissionais de campos de atuação distintos, especializados em áreas específicas da otimização de sítios. Acredita-se que a metodologia proposta nesta pesquisa viabiliza a estruturação de serviço inovador no mercado da informação.

Este documento é dividido em oito capítulos. Após introduzir o trabalho e o objetivo da pesquisa no presente capítulo, o Capítulo 2, para fins de contextualização, define os conceitos básicos encontrados na literatura clássica da ciência da informação e biblioteconomia sobre Sistemas de Recuperação da Informação; será explicado o que é Recuperação de Informação e o que são Sistemas de Recuperação de Informação; conceitos esses considerados pilares para a boa leitura e compreensão deste documento. O Capítulo 3 explica em detalhes como funcionam os motores de busca, considerando-os à luz dos seis

subsistemas apontados por Lancaster (LANCASTER 1993, pg. 15), detalhados no Capítulo 2. O Capítulo 4 descreve o processo de otimização de sítios e a sua diferença em relação às práticas mais conhecidas de *Spam*. No Capítulo 5, a metodologia proposta para realizar a otimização de sítios é exposta e comparada com a atividade de indexação tradicional. No Capítulo 6, descreve-se o experimento realizado aplicando a metodologia proposta assim como os resultados alcançados. O Capítulo 7 apresenta as conclusões da pesquisa. Finalmente, no Capítulo 8 são listadas as referências bibliográficas utilizadas.

2. Recuperação de Informação e Sistemas de Recuperação de Informação

Existem na literatura científica das áreas de ciência da informação, biblioteconomia e ciência da computação, inúmeras definições para o conceito Recuperação de Informação (RI). Meadow (1992, pg. 2), define RI como "encontrar alguma informação desejada em um repositório de informações ou banco de dados". Meadow ainda considera como sendo implícito nesta visão que a recuperação implica em seletividade, ou seja, uma atividade só pode ser considerada RI se houver a **seleção** de informação "que atenda a um conjunto de necessidades de informação altamente individualizadas". Já Lancaster (1993, pg. 11), define RI como "o processo de pesquisar uma coleção de documentos (usando o termo documento em seu sentido mais amplo) para identificar aqueles que tratem de um assunto em particular". Ambas as definições são muito semelhantes e, neste documento, a definição utilizada será a proposta por Lancaster.

Lancaster (1993, pg. 11) pondera se o termo RI seria o mais apropriado para descrever a atividade. Isto porque seu resultado real constitui-se de um ou vários documentos e não de informação: "... a transferência de informação só pode ocorrer se o usuário ler e entender o documento". Tal questionamento nos parece válido, mas não será aprofundado aqui. Atualmente o termo Recuperação de Informação é amplamente utilizado na literatura científica para designar o processo e será mantido neste documento. Uma discussão mais detalhada sobre o que é informação - que poderia ser utilizada numa discussão mais profunda e

abrangente sobre a impropriedade ou não do termo - é encontrada em Buckland (1991) ou em Cornelius (2002).

Grande parte do processo de RI é realizado na interação de um usuário e um Sistema de Recuperação de Informação (SRI). Porém, o processo de RI começa antes mesmo dessa interação ocorrer. Meadow (1992), divide o processo de RI em duas seqüências de ações: a **seqüência do usuário** e a **seqüência do produtor** do banco de dados (ou repositório de informações). Em seu texto, Meadow considera a fase: "Processamento da consulta do usuário pelo computador" como parte da **seqüência do usuário**. Neste documento, para ampliar o leque de SRI considerados, esta fase será interpretada como "Interação com o Sistema de Recuperação de Informação".

Sendo assim, as seguintes ações, identificadas por Meadow e brevemente discutidas aqui, são realizadas por um usuário no processo:

Percepção da ausência de alguma informação. Essa percepção pode ocorrer por vários motivos: um novo projeto, uma dúvida ocasionada durante a execução de uma tarefa, curiosidade, etc. Ainda segundo Meadow, "O reconhecimento de uma ausência de informação não significa necessariamente que o usuário saiba qual informação está ausente";

Expressão das necessidades. O próximo passo do usuário é tentar expressar quais são as suas necessidades. Este passo é essencial para todo o processo, é aqui que o usuário irá tomar consciência do que acha que precisa. Caso o resultado desta etapa seja ruim, é muito provável que o resultado de todo o processo seja ruim;

Elaboração de uma estratégia de busca. Neste passo, o usuário planeja como irá realizar a busca pela informação que necessita. O usuário pode realizar a busca indo direto a uma estante de uma biblioteca, olhando no catálogo de livros, procurando em algum diretório na internet ou ainda em um motor de busca;

Formulação da consulta. Neste passo, a necessidade de informação é colocada em termos apropriados para a utilização do Sistema de Recuperação de Informação. Isto pode ser tanto um conjunto de palavras-chave, no caso de uma consulta a um mecanismo de busca, quanto à seleção de índices, jornais e códigos de classificação, no caso de uma biblioteca, entre outras;

Interação com o Sistema de Recuperação de Informação. Neste passo, o usuário realiza a(s) consulta(s) ao SRI. O usuário pode, após esta fase, considerar o processo como terminado ou voltar a alguma fase anterior e recomeçar daquele ponto.

A seqüência do produtor do banco de dados é dividida em sete passos:

Decisão de criar o banco de dados. Este passo ocorre apenas uma vez no processo. Alguma organização verifica alguma possível ausência ou necessidade de fonte de informação e resolve criar um banco de dados para supri-la;

Decisão sobre o escopo do banco de dados. Após decidir pela criação do banco de dados é necessário definir qual o seu escopo. O banco de dados será geral? Ou será sobre um assunto específico (medicina, por exemplo)? Outro escopo a ser pensado nesta etapa é qual o tipo dos itens que serão representados: livros, revistas, arquivos digitais, quais formatos de arquivos, etc;

Decisão sobre a criação de registros individuais. Quais serão as informações de cada entidade que serão coletadas;

Seleção dos itens para inclusão. Quais serão os itens que serão disponibilizados. Por exemplo, em uma biblioteca poderia ser decidir se os artigos de uma revista serão incluídos ou somente informações sobre a própria revista. Num motor de busca poderia ser a decisão de quais páginas de um sítio seriam incluídas no banco de dados;

Criação do conteúdo dos registros individuais. Neste passo, os registros que serão incluídos na base de dados serão criados;

Entrada de dados. Inserção na base de dados dos registros criados;

Controle de qualidade. Este passo, que não ocorre após os outros, mas concorrentemente, tem como objetivo fazer com que os registros sejam produzidos e armazenados corretamente na base de dados.

Como dito anteriormente, a interação do usuário com um SRI é responsável por grande parte do processo de RI. Um SRI é definido por Meadow (1992, pg. 2) como "o sistema de computador, que consiste tanto de hardware como de software (...)". Este sistema seria o responsável pela Recuperação de Informação. Esta definição é limitada porque restringe a Recuperação de Informação à utilização de computadores. Uma definição mais ampla é dada por Salton (1983, prefácio xi): "...um sistema usado para armazenar itens de informação que precisam ser processados, pesquisados, recuperados e disseminados para várias populações de usuários". Uma definição similar - e que é a que será utilizada

neste documento - é dada por Lancaster (1993, pg. 11): "Qualquer sistema designado para facilitar esta atividade de busca de literatura pode legitimamente ser chamado de um Sistema de Recuperação de Informação".

Lancaster (1993, pg. 15), divide um SRI em seis componentes principais:

- Subsistema de seleção de documentos;
- Subsistema de indexação;
- Subsistema de vocabulário;
- Subsistema de procura;
- Subsistema de interface usuário/sistema;
- Subsistema de comparação.

É interessante notar que a maior parte das etapas do criador do banco de dados no processo de RI definido por Meadow são subsistemas de um SRI para Lancaster.

É apresentada a seguir uma descrição de cada subsistema.

Subsistema de seleção de documentos

Subsistema responsável por determinar quais documentos farão parte da base de dados. É composto pelas políticas e critérios de aquisição de novos documentos. Num ambiente como a *web* pode ser representado por um *crawler*, um programa que visita e armazena páginas ininterruptamente. Em uma biblioteca consistiria na política de aquisições de novos documentos, sejam estes livros, jornais, revistas, etc.

Subsistema de indexação

Após adquirir um documento, uma representação é criada para ele. A representação pode ser o seu assunto, as palavras-chave que o representam, etc. O subsistema de indexação é onde a representação do documento será criada. Esta representação será comparada com as representações das consultas dos usuários e, caso a similaridade entre ambas atinja algum critério de satisfação, a referência ao documento - ou o próprio documento – é "retornado" como resposta à consulta.

A indexação pode ser feita tanto manual quanto automaticamente. Por indexação manual entende-se que a criação da representação será feita por um ser humano, podendo ou não ter o auxílio de ferramentas. Na indexação automática, o trabalho será feito em grande parte, ou em sua totalidade, por máquinas. A indexação automática será melhor detalhada baseando-se em um motor de busca no Capítulo 3.

Subsistema de vocabulário

Vocabulário utilizado pelo sistema para representar o assunto dos documentos. Um SRI pode utilizar ou não um vocabulário controlado. Lancaster (1986) define um vocabulário controlado como "um conjunto limitado de termos que deve ser utilizado para representar o assunto de um documento". Segundo Svenonius (1989), a utilização ou não de um vocabulário controlado é uma das principais decisões na especificação de um SRI.

Ainda segundo Svenonius (1989), um sistema que não utiliza um vocabulário controlado pode ser caracterizado pela forma de indexação (linguagem natural, linguagem derivada, palavras-chave) e pela forma de busca permitida (texto completo ou texto livre).

Svenonius (1989) cita como motivo para a utilização de vocabulários controlados a suposição de que a tarefa de busca por informações é auxiliada por sua utilização. Isto por causa da crença de que sua utilização diminuiria o impacto de algumas características das línguas, como sinônimos e homônimos. Lancaster (1986) considera que os objetivos da utilização de um vocabulário controlado podem ser resumidos por:

- "Promover a representação consistente de assuntos tanto para indexadores quanto para usuários (...)";
- "Facilitar a condução de uma busca abrangente em algum tópico (...)".

Svenonius (1986) apresenta uma discussão interessante sobre a efetividade da utilização ou não de vocabulários controlados na recuperação de informação. Questões como o aumento da precisão e da revocação são tratadas nesse artigo.

Subsistema de interface usuário/sistema

Subsistema em que o usuário entra em contato com o SRI. Em uma biblioteca poderia consistir do profissional que conversaria com o usuário e procuraria entender qual é realmente a sua necessidade. Pode ser também o ambiente computacional em que um usuário entraria com uma consulta e receberia como resposta referências a documentos considerados pelo SRI como relevantes.

Subsistema de procura

Subsistema utilizado para tentar definir o que o usuário está procurando e idealizar uma estratégia de busca. Em um motor de busca, por exemplo, pode ser o mecanismo de refinamento de consultas; mais detalhes sobre este assunto podem ser encontrados no Capítulo 3.

Subsistema de comparação

Da mesma forma que um documento tem uma representação criada, uma representação também será criada para as consultas. Este subsistema tem como objetivo comparar as representações das consultas com as representações dos documentos.

O capítulo a seguir irá descrever sucintamente o funcionamento de um motor de busca utilizando para isso a subdivisão apresentada nesse capítulo.

3. Descrição de um motor de busca

Este capítulo é dividido em duas seções **Visão geral** (3.1) e **Subsistemas de um SRI em um motor de busca** (3.2). Na Seção **3.1** é realizada uma apresentação em alto nível de uma máquina de busca, enquanto na Seção **3.2** um motor de busca é detalhado à luz dos seis subsistemas apontados por Lancaster e apresentados no Capítulo 2.

3.1. Visão geral

Um motor de busca é uma classe de Sistema de Recuperação de Informação utilizada na internet que "modela a *web* como um banco de dados de texto completo (*full-text database*)" (BAEZA-YATES, 1999, pg. 373). Como exemplo de motores de busca pode-se citar o Google, Altavista ou Todobr[6]. Outro tipo de SRI muito utilizado na internet é o diretório: "taxonomia hierárquica que classifica o conhecimento humano" (BAEZA-YATES, 1999, pg. 385). Como exemplo de diretórios pode-se citar o Yahoo!, o Cadê?[7] ou o *Open Directory Project*[8]. Nos diretórios a classificação das páginas quanto à sua categoria é realizada por pessoas, contando com máquinas apenas como auxílio. Este documento não tratará de diretórios, focando somente em motores de busca.

Na visão de um usuário, um motor de busca é um sistema onde é entrada uma consulta que espera representar sua necessidade de informação e obtém como resposta um conjunto de URL's e, às vezes, outro conjunto de sugestões de novas

[6] http://www.todobr.com.br

[7] http://www.cade.com.br

[8] http://www.dmoz.org

consultas. Uma consulta pode ser formada por palavras que aparecem sozinhas ou por palavras e operadores, como, por exemplo, operadores booleanos AND, OR e NOT.

Procurando aprofundar o conhecimento sobre a natureza das consultas na *web*, Broder (2002) as divide em três tipos: consultas **navegacionais**, consultas **informacionais** e consultas **transacionais**.

- Consultas **navegacionais** são aquelas que têm como objetivo obter o endereço de um sítio que o usuário sabe ou assume que existe. Por exemplo, caso um usuário queira ir ao sítio do Ministério do Trabalho[9] e não saiba o seu endereço ele tem como alternativa acessar algum motor de busca e realizar a consulta +ministério do trabalho+; nesse caso, o usuário espera que entre as respostas retornadas conste o endereço do sítio do Ministério.

- Consultas **informacionais** são aquelas que têm como objetivo encontrar alguma informação já "pronta", não sendo necessário realizar qualquer outra interação com o recuso encontrado além de acessá-lo e lê-lo. Como exemplo desse tipo de consulta pode-se supor um usuário que esteja fazendo uma pesquisa a respeito de todos presidentes brasileiros e realizaria, num motor de busca, a consulta +presidentes do Brasil+ esperando obter como resposta algum endereço com tal informação.

[9] http://www.mte.gov.br

- Consultas **transacionais** são aquelas que tem como objetivo encontrar um determinado tipo de sítio no qual outras interações ocorrerão. Por exemplo, encontrar algum sítio que venda livros.

Baseando-se nessa classificação, Broder apresenta também uma divisão da evolução dos motores de busca em três gerações.

- Primeira geração: Tem suporte principalmente a buscas informacionais. Utiliza basicamente dados da própria página e dados como o número de páginas da coleção em que um termo da consulta aparece. Era muito similar ao modelo tradicional de recuperação automática de informação.
- Segunda geração: Utilização de maior quantidade de dados de fora da página e específicos da *web*, como, por exemplo, a análise de apontadores ou o texto de âncora. As consultas navegacionais são contempladas nesta geração.
- Terceira geração: Tem como objetivo conseguir descobrir mais dados sobre o contexto da busca, iniciando pelo tipo da consulta (navegacional, informacional ou transacional) e exibir melhores resultados ao usuário.

É importante frisar que a passagem dos motores de busca de uma geração para outra não significou a substituição das técnicas da geração anterior pelas da nova, mas sim uma adição de novas técnicas àquelas existentes.

3.2. Subsistemas de um SRI em um motor de busca

3.2.1. Subsistema de seleção de documentos

Este subsistema é representado nos motores de busca pelo *Crawler*. Um *crawler* é um programa que executa continuamente, visitando páginas que poderão ser adicionadas posteriormente ao índice do motor.

Visto de forma bem simples, um *crawler* funciona da seguinte forma: Existe uma fila inicial de URL's a serem visitadas. Cada URL é visitada e os apontadores encontrados na página na URL são colocados na fila para serem visitados depois. Existem motores que permitem que usuários coloquem URL's nessa fila, já outras não, por considerarem que as URL's adicionadas são, provavelmente, de pouco valor.

As requisições feitas por um crawler a um servidor *web* podem sobrecarregá-lo (o servidor), por isso um guia com regras básicas sobre como as requisições devem ser feitas foi definido (KOSTER, 1993). Todo *crawler* deve seguir estas regras. Além disso, o responsável por um sítio pode desejar que partes deste não sejam coletadas, para garantir que isso aconteça, um arquivo (robots.txt) especificando o que não deve coletado ser colocado na raiz do sítio.

São diversas as formas que um *crawler* pode funcionar. Existem *crawlers* que só coletam páginas de uma determinada língua ou país (em motores de busca verticais; o Todobr é um exemplo) enquanto que outros coletam páginas em qualquer idioma. A ordem com que páginas são coletadas também é levada em consideração. Existem estudos (CHO, 1998) que discutem como coletar páginas

de maior qualidade primeiro. Uma breve discussão a respeito da atividade do *crawler* pode ser encontrada em (BAEZA-YATES, 1999, pg. 382).

3.2.2. Subsistema de indexação

É o subsistema onde a criação de índices é realizada automaticamente. A forma como os índices são criados varia de motor para motor, aqui exporemos algumas técnicas já divulgadas por alguns motores (mesmo que sem detalhes).

Como já foi exposto, existem três gerações de motores de busca. As técnicas utilizadas na primeira geração eram praticamente as mesmas utilizadas na indexação automática tradicional, ou seja, baseavam-se em grande parte no Modelo de Espaço Vetorial proposto por Salton (1975). Explicando bem sucintamente, este modelo considera cada consulta como um documento e compara a sua similaridade com cada documento da coleção. Um documento é representado como um vetor n-dimensional, onde n é o número de termos da coleção. A similaridade entre dois documentos é dada pelo cosseno do ângulo entre ambos, ou seja, quanto menor o ângulo entre os vetores, maior a similaridade. Em uma consulta os termos que a constituem não possuem a mesma importância, a cada termo é atribuído um peso; o valor de cada termo no cálculo da similaridade é dado pelas medidas `Tf` (*term frequency* – freqüência do termo) e `Idf` (*inverse document frequency* – inverso da freqüência de documentos). O `Tf` é uma medida relacionada ao número de vezes que um determinado termo aparece em um documento - quanto mais um termo aparece em um documento maior o `Tf` desse termo nesse documento. O `Idf` é uma medida da raridade do termo na coleção - quanto maior o número de documentos

da coleção em que o termo aparece, menor seu Idf na coleção. O raciocínio por trás destas medidas pode ser ilustrado da seguinte forma. Suponha uma consulta com dois termos, *Ta* e *Tb*. Se o número de documentos em que *Ta* aparece for menor do que o número de documentos em que *Tb* aparece, isso significa que *Ta* é um termo de maior peso do que *Tb* pela medida do Idf (mais raro e, portanto, melhor discriminante). Da mesma forma, se em um documento o termo *Ta* aparece mais em seu corpo do que nos demais da coleção para o termo *Ta* esse documento é considerado como mais relevante do que os outros da coleção. O peso de um termo em cada documento é dado pelo produto Tf x Idf. A relevância atribuída à cada documento da coleção é computada através da soma do peso encontrado para cada termo da consulta no documento – calculado através do produto Tf x Idf – e de uma posterior normalização do valor encontrado. Essa técnica foi, e ainda é, utilizada nos motores de busca, mas com pequenas alterações. Por exemplo, caso um termo ocorra no corpo da página tem um peso, caso ocorra no título tem outro.

Na segunda geração, dados específicos da *web* como a análise de apontadores e o número de visitas a uma página foram adicionados. O primeiro motor a utilizar a análise de apontadores foi o Google. Seu algoritmo, o Pagerank (PAGE, 1998), considera que cada apontador de uma página para outra é um voto. Quantos mais votos uma página possuir, mais importante ela é considerada. Esta análise é chamada de análise global, já que é feita independentemente de uma consulta em particular. Kleinberg (1998), desenvolveu um algoritmo que realiza a análise de apontadores no resultado de uma consulta, ao invés da coleção. Esta análise é

chamada de análise local. Outros dados, como o texto de âncora dos apontadores, também são utilizados nessa geração.

Na terceira geração, dados sobre o contexto da consulta começam a ser utilizados para a construção dos índices. Dados sobre a página, como o país onde esta se localiza ou o idioma em que seu texto está redigido começam a ser incorporados.

3.2.3. Subsistema de vocabulário

O vocabulário utilizado em um motor de busca é estritamente aquele encontrado nos documentos da sua coleção. As atividades mais próximas do controle de vocabulário são o sistema que elimina palavras que não acrescentam qualquer valor a uma consulta - por aparecer em praticamente todos documentos da coleção, como, por exemplo, "de", "e", "o", em português, ou "the", em inglês - e a atividade de extração de radicais, para permitir que consultas como, por exemplo, +menino+ e +meninos+ tenham o mesmo resultado.

3.2.4. Subsistema de procura

Este subsistema foi incorporado nos motores de terceira geração e é caracterizado por ferramentas que tentam esclarecer qual o contexto da consulta, o que o usuário realmente precisa. Um exemplo é a sugestão de novas consultas baseadas no histórico de consultas de um usuário em particular, dos usuários do motor em geral e na análise de páginas em que os termos da consulta original aparecem. Anick (2003) relata um estudo realizado a respeito no motor Altavista.

Como exemplo de motores em que este conceito está sendo utilizado extensivamente são o Teoma[10] e Vivísimo[11].

3.2.5. Subsistema de interface usuário/sistema

Nos motores de busca, a interface geralmente consistia em uma página em que o usuário entra com a consulta e recebe como resposta o conjunto de URL's para páginas consideradas relevantes. Nos motores de busca atuais o conjunto de respostas é acompanhado por sugestões de novas consultas.

3.2.6. Subsistema de comparação

Compara a representação da consulta de um usuário com o índice do motor. O resultado dessa comparação é uma lista de referências a documentos; as referências são apresentadas em ordem decrescente da relevância atribuída a cada documento.

O processamento gasto nesta etapa é um dos aspectos observados por um usuário na formação de sua opinião sobre o motor. Caso o motor demore muito para responder o usuário deixará de utilizá-la. A importância da velocidade é tão grande que, em 2003, o Google utilizava cerca de 15000 computadores distribuídos pelo mundo para o processamento de consultas (BARROSO, 2003).

Algumas tentativas de melhorar a relevância do resultado de uma consulta ocorrem neste subsistema também. A já citada análise local de apontadores ocorre nesta etapa. Outro método que ocorre nesta etapa é a expansão automática de consultas. Este método consiste em modificar a consulta dos

[10] http://www.teoma.com

[11] http://www.vivisimo.com

usuários aumentando o número original de termos sem que ele tenha consciência disso.

4. Otimização de sítios *web*

Este capítulo descreve em detalhes o processo de otimização de sítios (Seção 4.1) e o que o diferencia da prática de *Spam* (Seção 4.2).

4.1. Descrição da atividade de otimização

Otimização de sítios é o nome dado para a atividade de adequar uma página HTML (*HyperText Markup Language*), ou um conjunto de páginas, a fim de alcançar melhores posições para determinadas consultas em um motor de busca.

A otimização de sítios é uma atividade que interessa a qualquer organização que queira divulgar algo na Internet e queira ser encontrada através de algum motor de busca. Por exemplo, suponha um sítio de divulgação de notícias médicas. Caso esse sítio apareça nas dez primeiras posições para a consulta "notícias médicas" em um grande motor de busca, é muito provável que receba muitas novas visitas.

A otimização de sítios consiste em fazer com que uma página possua o maior número possível das evidências que um motor de busca utiliza para determinar sua relevância. Porém, como já foi dito, nem todas as evidências utilizadas, nem os detalhes a respeito daquelas conhecidas, são divulgados. Este fato, entretanto, não impede que a atividade seja realizada. Nem todos os detalhes precisam ser conhecidos e esses, na maior parte das vezes, interessam apenas aos próprios motores. Além disso, as evidências não conhecidas, mas que influenciam no cálculo da relevância, podem ser "deduzidas" observando-se os primeiros resultados para grupos de consultas. Como se vê, tal ambiente não se presta à estruturação de uma atividade de pesquisa determinista no estrito senso.

Entretanto a dificuldade na obtenção de resultados deterministas não precisa ser considerada um obstáculo instransponível para o processo de otimização de sítios. Como mostra esta pesquisa, bons resultados podem ser alcançados.

O processo de otimização de sítios começa com a determinação das palavras-chave interessantes para uma página - aquelas que descrevam o seu conteúdo e pelas quais o responsável pela página quer que seja encontrada. Essa etapa é crucial para o sucesso do processo como um todo. Uma página deve estar, preferencialmente, na primeira página de respostas para todas as consultas com palavras-chave que um "usuário de interesse[12]" procure. Por usuário de interesse, entende-se um usuário que esteja procurando por alguma informação relacionada àquela disponibilizada na página. Essa etapa envolve a compreensão do assunto da página, o conhecimento do público para o qual ela foi concebida e as palavras-chave utilizadas por este público em suas buscas.

A outra etapa consiste em posicionar as palavras-chave no maior número possível de locais de evidências, tanto evidências internas quanto externas.

O posicionamento relativo às evidências internas é aparentemente simples, bastaria que se conhecessem os locais em que os motores de busca as procuram. Porém, nem todos os casos são simples. Deve-se lembrar, sempre, que páginas são feitas para serem lidas por pessoas, não por motores, ou agentes de software. Distribuir aleatoriamente as palavras pelo texto (o número de ocorrências de palavras no texto é uma das evidências mais conhecidas) pode até ser efetivo do ponto de vista de motores, mas não é adequado para um leitor, que pode

[12] Usuário integrante do público alvo focalizado pelo sítio.

desconsiderar o conteúdo da página. Isso exemplifica a necessidade de se trabalhar bem a forma de posicionar as palavras nos locais de evidências que contêm informações úteis para o usuário de interesse (a maioria). Nos locais onde não se veicula informação útil para o usuário - tais como nomes de domínios, subdomínios, diretórios, arquivos e *meta-tags* dentro dos arquivos -, as palavras podem ser colocadas sem muito trabalho, embora devam ser escolhidas com cuidado, pensando sempre naquelas palavras que terão maior probabilidade de serem utilizadas pelo público alvo do sítio.

O posicionamento de evidências externas é mais complicado do que o de evidências internas. O posicionamento de evidências em outras páginas do mesmo sítio, por exemplo, nas âncoras dos apontadores para a página sendo otimizada, é feito da mesma forma que o posicionamento de evidências internas. A respeito das evidências externas que não estão sobre controle do responsável pela página ou pelo sítio, como, por exemplo, obter apontadores de páginas em outros sítios, não há muito o que ser feito, ainda que seja possível tentar estabelecer parcerias com outras instituições não concorrentes, para a troca de apontadores entre seus sítios.

4.2. Otimização de sítios x *Spam*

Spam é uma prática muitas vezes confundida com a otimização de sítios por ambas compartilharem o mesmo objetivo: fazer com que uma página apareça na lista das dez primeiras URLs para um conjunto de consultas. A diferença entre essas duas abordagens nem sempre é muito clara à primeira vista. Henzinger (2002), define *spam* como o processo de "deliberadamente tentar manipular sua

colocação no *ranking* de vários motores de buscas"[13]. Essa é uma definição com a qual não se concorda neste documento, já que também na otimização tenta-se manipular a colocação de uma página no ranking. Porém, ao contrário do que ocorre na prática de *spam*, a otimização de sítios não lança mão de táticas que objetivam enganar os motores. A conceituação que será utilizada neste documento é a de que spam é qualquer técnica que tenha a intenção de enganar um motor e fazer com que a uma página seja atribuída uma relevância superior àquela que ela realmente tem.

Ainda segundo Hezinger, as técnicas de *spam* podem ser divididas em três grandes grupos: *Text Spam, Link Spam e Cloaking*.

Text Spam é o conjunto de técnicas que se baseiam na modificação do texto da página de forma que o motor a atribua uma relevância maior do que a percebida pelo usuário. Um exemplo da prática de *Text Spam* é quando inúmeras palavras-chave interessantes são colocadas com a cor da fonte da mesma cor do fundo da página; dessa forma, um usuário não percebe que o texto está na página, porém um motor pode vir a processá-lo e atribuir para a página, para consultas por aquelas palavras escondidas, uma relevância que a mesma não deveria ter.

Conforme explicado na Seção **3.1**, os motores de busca, a partir da segunda geração, começaram a utilizar dados específicos referentes ao contexto da página na *web* para tentar inferir a relevância de uma página; os apontadores são um bom exemplo. *Link Spam* é o nome dado para a atividade de tentar manipular a

[13] Tradução arbitrária para "*deliberately manipulate their placement in the rankings of various search engine*"

análise de apontadores realizada pelos motores. Um exemplo de *Link Spam* é a construção de um grande número de páginas sem qualquer conteúdo além de apontadores para uma determinada página; o benefício que tal página obteria seria um grande número de apontadores, além dos textos de âncora de cada apontador.

O último grupo de técnicas é o chamado *Cloaking* (de *Cloak*, disfarçar, cobrir, ocultar). A idéia por trás das técnicas desse grupo é entregar ao *crawler* um conteúdo totalmente diferente daquele entregado aos usuários. Por exemplo, suponha que o *designer* uma página pornográfica acredite que o grupo das pessoas que realizam a consulta +ronaldinho+ seja formado por potenciais clientes. Como o conteúdo do sítio não tem como ser modificado para que suas páginas obtenham uma boa posição para essa consulta, a alternativa encontrada é criar uma página otimizada para +ronaldinho+ que será entregue para os *crawlers* e uma página com pornografia que será entregue para as pessoas.

Como resposta a esses esforços de manipulação, as empresas responsáveis pelos motores desenvolveram e, continuam a pesquisar e desenvolver, métodos para a detecção de *Spam*. Apesar dos motores não divulgarem suas técnicas, existem estudos na literatura científica que podem ser consultados. Fetterly (2004) propõe um conjunto de técnicas baseadas em análise estatística para a detecção de *Spam* em páginas geradas automaticamente com essa finalidade; algumas propriedades de conectividade na *web* são utilizadas por Amitay (2003) para a detecção de *link spammers*. Uma revisão mais detalhada de literatura sobre *Spam* e sua detecção foge do escopo desta dissertação, porém leitores interessados no

assunto podem consultar, além dos artigos anteriormente citados, os trabalhos de Davidson (2001) e Dwork (2001).

Talvez o aspecto mais importante dessa discussão seja o fato de que, caso algum sítio seja surpreendido por algum motor praticando *Spam,* este certamente estará infringindo suas regras de utilização e sofrerá penalidades. Tais penalidades podem variar desde a perda de algumas posições no *ranking,* até a retirada de todas as páginas do sítio e subseqüente não indexação do mesmo.

5. Metodologia para a otimização de sítios para motores de busca

Esta pesquisa divide o processo de otimização de sítios em duas grandes etapas:

- Etapa A: Fazer com que a página seja indexada;
- Etapa B: Fazer com que a página apareça para consultas de interesse.

Essas etapas serão tratadas neste capítulo em duas seções. A primeira discorre sobre como fazer para uma página ser indexada por um motor de busca. A segunda explica como a otimização pode ser compreendida como uma atividade de indexação, utilizando-se dos planos de Ranganathan.

5.1. Fazendo com que uma página seja indexada

O subsistema de seleção e coleta de documentos é essencial para o processo de otimização de sítios por um motivo evidente: páginas não coletadas não são processadas nem incluídas no índice do motor de busca, conseqüentemente elas não aparecem nos resultados de consultas.

Um "otimizador" de sítios deve considerar os seguintes aspectos ao pensar em uma estratégia para sua página ser coletada pelo mecanismo:

- Páginas geradas dinamicamente;
- Submissão de sítios;
- Obtenção de apontadores.

5.1.1. Páginas geradas dinamicamente

Na internet um número cada vez maior de páginas são criadas no momento em que são requisitadas, são páginas geradas dinamicamente. Essas páginas podem

ser criadas, por exemplo, a partir de consultas a banco de dados ou do resultado da execução de um programa. Exemplo de páginas geradas dinamicamente são as páginas que contém os resultados de consultas a qualquer motor de busca. Nem todos os motores coletam páginas geradas dinamicamente e mesmo aqueles que coletam não o fazem com a mesma freqüência com que coletam páginas estáticas - aquelas que já estão prontas no servidor, não são resultado de nenhum processamento em especial.

Para diferenciar uma página gerada dinamicamente de uma página estática, o *crawler* observa se o caractere '?' aparece na URL que será requisitada; caso o caractere ocorra, é certo que a página é gerada dinamicamente (o motivo para essa certeza será explicado posteriormente), caso não ocorra, não há como fazer tal diferenciação[14]. O motivo para tal certeza é a definição de que os parâmetros para o programa que irá processar a requisição vêm sempre após esse caractere. Isso segundo o padrão de funcionamento de URLs na *web*. Por exemplo, suponha que um usuário resolva pesquisar no motor MEUMOTOR e a URL dessa consulta seja:

http://www.MEUMOTOR.com.br/busca?termos=rede+de+computadores

Examinando a URL, e sabendo como a passagem de parâmetros é codificada, determina-se que a consulta enviada para o motor é +rede de computadores+.

[14] Nesse caso a página pode ser estática ou dinâmica, o *crawler* não dispõe de nenhum indício do fato para diferenciar os dois casos.

Para evitar que uma página não seja coletada por um *crawler* por ser gerada dinamicamente é necessário lançar mão de alguma técnica que faça com que seja considerada estática.

Uma técnica para tanto é a **reescrita de URL**. O funcionamento básico da técnica é a transformação interna (no servidor *web*) de uma URL que pareça ser de um recurso estático para uma URL de um recurso gerado dinamicamente. Por exemplo, suponha as URLs

(1) http://servidor/programa?arg1=valor1&arg2=valor2 e
(2) http://servidor/programa?arg1=valor1.1&arg2=valor2.1

É claro para um *crawler* que ambas referenciam recursos gerados dinamicamente. Já as URLs:

(1') http://servidor/programa/arg1/valor1/arg2/valor2 e
(2') http://servidor/programa/arg1/valor1.1/arg2/valor2.1

não podem ser consideradas como sendo geradas dinamicamente.

O funcionamento da técnica de reescrita de URL é fazer com que (1') e (2') sejam as URLs encontradas pelos *crawlers* e usuários do sítio, mas o programa que irá processar os parâmetros receba (1) e (2). Os detalhes técnicos de como isso pode ser feito fogem ao escopo deste documento podendo, porém, serem encontrados em diversos documentos que descrevem servidores *web* e suas funcionalidades, como, por exemplo, *Apache rewrite guide*[15], e *URLRewriting IIS*[16].

[15] http://httpd.apache.org/docs2.0/misc/rewriteguide.html

[16] http://msdn.microsoft.com/library/en-us/dnaspp/html/URLRewriting.asp

5.1.2. Considerações sobre a submissão de sítios

Tradicionalmente, todos os motores aceitavam a submissão de URL's. Porém, com o passar do tempo, inúmeros programas foram criados para realizar a submissão automática e contínua de páginas em diversas motores (alguns desses programas alardeiam submeter uma página em mais de 5000 motores!) - por submissão contínua, entenda-se que uma mesma URL é periodicamente re-submetida a um motor. Tal submissão descontrolada fez com que a fila de URL's obtidas pelas submissões de usuários fosse composta, em sua maioria, por URL's de páginas já indexadas ou de baixa qualidade, com pouco ou nenhum conteúdo informacional relevante para qualquer consulta.

Para evitar que seus *crawlers* despendessem recursos coletando páginas inúteis, os motores resolveram modificar sua política concernente à submissão de URL's. Alguns, como o TodoBr[17], decidiram não aceitar mais submissões de URL's. Isso significa que uma página será coletada somente se alguma outra página já no índice apontar para ela. Outros, como o Altavista, criaram mecanismos que garantiram que a submissão gratuita de URL's seja feita apenas por pessoas, enquanto que a submissão automática de um número determinado de URL's seja paga. Ao que parece essa modalidade (submissão automática paga) foi extinta com a aquisição do Altavista pelo Yahoo!. Os motores que optaram por continuar a aceitar a submissão indiscriminada de URL's, como o Google, parecem não dar muita atenção para as URL's submetidas.

[17] http://www.todobr.com.br/faqs.html

A submissão de um sítio em motores de busca não é hoje em dia um meio efetivo de conseguir a inclusão nos índices, devendo ser feito somente por "desencargo de consciência". A maneira mais efetiva de ser adicionado ao índice de um motor de busca é através da obtenção de apontadores de páginas que já estejam indexadas. Tal método será discutido na subseção seguinte.

5.1.3. Obtenção de apontadores

A forma mais utilizada por motores de busca para a adição de novas páginas ao seu índice é seguindo apontadores que estejam em páginas já visitadas pelo seu *crawler* e indexadas, o que demonstra a importância de se obter apontadores para as páginas de um sítio sendo otimizado. Além disso, os motores atuais quando calculam a relevância de uma página para uma consulta levam em consideração a sua "popularidade" (da página), que, por sua vez, é inferida através do processamento dos apontadores encontrados.

Porém, obter apontadores não é uma tarefa simples. Um apontador de um sítio para outro pode ser pensado como uma indicação de que o sítio apontado possui informações relevantes para os usuários daquele que aponta; essa interpretação implica que para um sítio obter um apontador ele necessariamente deve possuir um conteúdo digno de ser referenciado. Construir páginas com conteúdo relevante foge do escopo do trabalho de um otimizador, que deve, contudo, pensar em como fazer para auxiliar a obtenção.

Uma forma de obter apontadores é submeter a página principal do sítio para avaliação em um diretório, como o Yahoo! ou Open Directory; caso o avaliador, uma pessoa, considere que a página seja pertinente para alguma das categorias

do diretório então o endereço do sítio será colocado dentro dessa categoria. Deve-se lembrar que em diretórios como o Yahoo!, atualmente, é cobrada uma taxa para a avaliação de sítios com finalidade comercial.

Um otimizador, mesmo que não seja o responsável pela criação do conteúdo de um sítio, deve atuar ativamente para a obtenção de apontadores. Isso pode ser feito através da busca de sítios que possam obter algum benefício apontando para páginas do sítio sendo otimizado, do aconselhamento dos responsáveis sobre a importância de fazer um trabalho de divulgação do sítio, lembrando sempre a importância da criação de conteúdo relevante e, evidentemente, alertando sobre os perigos de obter apontadores de forma ilícita, através de técnicas de spam (cf. Seção **4.2**). O otimizador deve também alertar os responsáveis pelo sítio para a importância do estabelecimento de parcerias com outros sítios complementares, com o propósito de troca de apontadores.

5.2. Fazendo com que a página apareça para consultas de interesse

5.2.1. Planos de Ranganathan

Nesta seção, o processo de indexação manual de documentos será brevemente apresentado com o objetivo de dar um embasamento para o paralelo entre esta e o processo de otimização de sítios, que será traçado na próxima subseção.

O objetivo da indexação de documentos segundo Lancaster (1991) é "construir representações de um documento numa forma que se preste à sua inclusão em algum tipo de base de dados".

Ranganathan (1962), divide a atividade em três planos de trabalho:

1. Plano Idéia: análise do assunto do documento;
2. Plano Verbal: exame de vocabulários controlados, tesauros, das tabelas e dos índices do esquema de classificação para encontrar os conceitos identificados no primeiro plano;
3. Plano Notacional: construção da notação para os conceitos de acordo com as regras estabelecidas.

O Plano Idéia tem como objetivo descrever o assunto sobre o qual o documento trata. Neste plano, a questão de como o assunto do documento será representado não deve ser considerada, esta tarefa será realizada nos planos seguintes. O resultado deste plano é uma frase, nas palavras do indexador, descrevendo o documento considerado.

O Plano Verbal é o plano em que os conceitos encontrados no Plano Idéia (representado por meio de uma frase nas palavras de quem fez a análise) serão traduzidos para um conjunto de conceitos. Por exemplo, suponha que a análise de um documento produziu a frase "O documento trata sobre engenharia naval e engenharia mecânica no Brasil". Após consultar um vocabulário controlado, o indexador irá obter os conceitos "Engenharia naval", "Engenharia mecânica" e "Brasil". O resultado deste plano é o conjunto de conceitos que representam o assunto encontrado no primeiro plano.

O Plano Notacional é o plano em que a notação para o documento é construída, utilizando-se as representações dos conceitos encontrados no plano verbal em um esquema e as regras definidas pelo esquema. Por exemplo, suponha os conceitos

"Engenharia naval", "Engenharia mecânica" e "Brasil" encontrados no plano verbal. Suponha também que o indexador esteja utilizando a CDU[18]. Após examinar as tabelas e índice da CDU o indexador encontra os códigos 623, 621 e 81, representando os conceitos. De acordo com as regras de construção de notação da CDU, cuja explicação foge ao escopo deste documento, estes códigos seriam agrupados e formariam a seguinte notação: [621+623](81). O resultado deste plano é a representação final do assunto de um documento.

5.2.2. Utilização dos planos de Ranganathan na otimização de sítios

Entre as atividades de indexação tradicional e de otimização de sítios, uma diferença se nota imediatamente: enquanto o resultado da indexação tradicional de um documento é uma notação ou um conjunto de termos, o resultado da otimização é um novo documento. Além disso, o ambiente em que o resultado da indexação será utilizado é controlado por aquele que está disponibilizando o documento, enquanto que na otimização, quem disponibiliza não tem qualquer controle sobre o ambiente.

Porém, existem semelhanças entre ambas atividades, que permitem que a otimização de sítios seja considerada à luz de uma atividade de indexação tradicional, apropriando-se de conhecimentos clássicos e já consolidados da área de biblioteconomia e ciência da informação. Com efeito, o objetivo de ambas as atividades é o mesmo, qual seja: facilitar para um usuário a tarefa de encontrar um documento que esteja relacionado com sua necessidade de informação. Outra

[18] Classificação Decimal Universal. Esquema de classificação utilizado para auxiliar a tarefa de indexação manual de documentos.

semelhança é que ambas as atividades são feitas por especialistas humanos, contando com auxílio de motores e outros agentes de software apenas para as tarefas mais repetitivas, não para as intelectuais.

Verificou-se nesta pesquisa que a semelhança entre as duas atividades permite que se defina a atividade de otimização de sítios como um tipo especial de indexação, abrindo assim a possibilidade de utilizar os três planos de Ranganathan para estruturar a otimização de uma página, beneficiando-se assim de alguns conceitos já consolidados originários da biblioteconomia.

Adiante é descrito como os três planos poderiam ser utilizados:

1. **Plano Idéia**: análise da página para descobrir o seu assunto. Caso exista mais do que um assunto, verificar se a página não pode ser dividida em mais páginas. Nota-se que a atividade, neste plano, na otimização, é muito similar àquela feita na indexação tradicional;
2. **Plano Verbal**: descobrir quais são as palavras que melhor descrevem o assunto da página. Na atividade de otimização não existem vocabulários controlados para auxiliar a escolha de palavras-chave (descritores), já que o ambiente *web* não possui qualquer controle. O que deve ser feito então é usar um conjunto inicial de palavras e verificar, através de algumas ferramentas, se os usuários estão utilizando-as para buscas nos motores de busca. Esta atividade deve ser repetida até ser encontrado um conjunto satisfatório de palavras. Nem sempre as palavras encontradas nesta fase serão realmente as que os usuários utilizarão nas suas buscas. Neste caso, deve-se monitorar por quais palavras os usuários estão chegando à página

otimizada e refazer esta etapa. Em paralelo é necessário registrar o `Idf` de cada palavra do conjunto, e mantê-las ordenadas decrescentemente;

3. **Plano Notacional**: dado que o otimizador possui as palavras, resta "apenas" posicioná-las nos locais de evidências internas e externas. Uma discussão a respeito deste posicionamento, assim como a descoberta dos locais, será feita na seção "Detalhamento da utilização do plano notacional" (Seção 5.2.3).

Além de possibilitar a utilização de conhecimentos da indexação tradicional em um novo contexto, a estruturação do trabalho em planos possibilita a divisão em sub-tarefas que podem ser repassados a especialistas distintos. Por exemplo, o profissional responsável pelo Plano Idéia não precisa conhecer a fundo como os mecanismos de busca funcionam e esta tarefa ficaria a cargo do responsável pelo Plano Notacional. Um esboço da divisão de trabalhos pode ser visto na tabela abaixo:

Plano	Conhecimento necessário
Idéia	Análise de assuntos. Às vezes, de mais de um tipo de formato: texto, imagem, vídeo etc.
Verbal	Conhecimento de ferramentas e métodos para descobrir por quais palavras os usuários de interesse (que são o público alvo do sítio) estão pesquisando.
Notacional	Conhecimento de como os motores de busca funcionam, quais as evidências que estão sendo procuradas. Habilidade de posicionar as palavras nos locais de evidências sem tornar a página desagradável para o usuário.

Tabela 1 Conhecimento necessário para as etapas de otimização.

A subseção seguinte provê um detalhamento da utilização do plano notacional na otimização de sítios.

5.2.3. Detalhamento da utilização do plano notacional

Como visto na Seção 3.2, no modelo vetorial de indexação proposto por Salton (1975), existem duas medidas utilizadas para o cálculo da relevância de uma página em relação à uma dada consulta, o `Tf`[19] e o `Idf`[20]. Uma estratégia direta para aumentar a relevância atribuída a uma página para uma consulta é aumentar a freqüência dos termos da consulta no corpo da página. Por exemplo, suponha a página com o seguinte texto:

[19] Número de vezes que as palavras da consulta aparecem na página.

[20] Que valoriza as palavras mais raras da consulta, que são consideradas melhores discriminantes.

O menino jogava futebol alegremente com seus colegas. Isso foi tudo que ele sempre gostou de fazer. Quando tinha 5 anos ele foi presenteado com uma bola por sua mãe, desde então seu passatempo é chutá-la.

Uma forma de otimizá-la para a consulta +bola de futebol+ seria reescrevê-la como:

O menino jogava futebol alegremente com seus colegas. Jogar bola foi tudo que ele sempre gostou de fazer. Aos 5 anos ele ganhou uma bola de futebol de sua mãe, desde então seu passatempo é chutá-la.

Os números de aparições dos termos da consulta para a página não otimizada e otimizada podem ser vistos na tabela abaixo:

	Não otimizada	Otimizada
Bola	1	2
De	1	3
Futebol	1	2

Tabela 2 Número de aparições de termos em um texto otimizado e outro não otimizado.

Para a consulta referida, o `Tf` de cada termo na página otimizada seria maior e, portanto, a relevância atribuída à página também seria maior. Ou seja, a página otimizada apareceria em melhor posição que a página não otimizada.

É importante notar que apenas o fato dos termos da consulta aparecerem mais vezes na página otimizada não garante que essa alcançará melhor posição. Isso ocorre por dois motivos: o modelo vetorial não é o único método utilizado para avaliar a relevância de uma página e, mesmo que fosse, o número de palavras que uma página contém também influencia o resultado. No exemplo acima, as

duas páginas possuem o mesmo número de palavras e, por isso, a estratégia funcionou.

No caso das páginas possuírem um número diferente de palavras ou caso alguns termos da consulta apareçam mais em uma página enquanto os demais apareçam mais na outra, não é tão simples determinar qual das páginas é mais relevante. Para tanto, seria necessário conhecer o `Idf` dos termos da consulta e qual a fórmula exata utilizada pela máquina para o cálculo da relevância – o que não é divulgado, por razões obvias.

Ainda assim, é possível delinear três diretrizes para a otimização de sítios com base exclusivamente no conhecimento do modelo vetorial:

- Repita o maior número de vezes possível no corpo da página os termos da consulta para qual a otimização está sendo realizada;
- Quando, por algum motivo, deva-se fazer uma decisão sobre qual termo utilizar em determinado posicionamento dê preferência para os termos com maior `Idf`;
- Evite fazer páginas muito grandes, que tratem de vários assuntos diferentes. Sempre que possível divida páginas com tais características em várias outras, cada uma tratando de um assunto bem definido. Com isso, além de cada página também ter um conjunto de palavras-chave de interesse bem definido, sendo possível focar o trabalho de otimização, evita-se que os documentos sejam penalizados pela normalização feita no Modelo Vetorial.

Para determinar os termos com maior `Idf`, basta, para cada termo, realizar uma busca no motor para qual a página está sendo otimizada e verificar o número de URLs retornadas como resultado. Quanto menor o número de URLs retornadas como resultado, maior o `Idf` do termo.

Uma aplicação direta das diretrizes acima leva ao seguinte raciocínio: para a página aparecer bem para uma dada consulta, basta então repetir os termos da consulta arbitrariamente no corpo da página. Ou então, pode-se criar uma página, completamente sem sentido aparente, composta apenas pelos termos da consulta, que direciona o usuário a outra página em que os termos não apareçam. Essas são duas práticas conhecidas do que é chamado de *spam* em motores de busca. Caso um motor de busca identifique uma página recorrendo a tais métodos, esta última será duramente punida pela primeira. Considerações a respeito dessas práticas, e outras similares, são feitas na Seção 4.2.

Tais punições levam naturalmente ao refinamento da primeira diretriz para:

- No corpo da página, repita os termos da consulta sempre que possível, preservando a semântica original do conteúdo;

Caso a otimização seja feita de acordo com essa diretriz, o conteúdo retornado ainda será relevante para um usuário da máquina de busca não havendo, portanto, motivos para a página ser considerada *spam* e punida.

Como dito na Seção 3.2, os termos que aparecem em diferentes partes da página, por exemplo, no título ou no primeiro parágrafo, podem ter pesos diferentes atribuídos no cálculo da relevância; mais do que isso, termos que aparecem no

texto de âncora de um apontador também influenciam na relevância atribuída à página apontada. Mesmo nesses casos as diretrizes continuam válidas, sendo importante lembrar que o texto de âncora do apontador para a página só pode ser modificado pelo responsável pela página que aponta.

6. Experimento

Neste capítulo, apresenta-se o experimento realizado visando à verificação empírica da eficácia da metodologia proposta nesta pesquisa. O experimento é de natureza qualitativa; trabalhos futuros serão necessários para a realização de um estudo mais profundo e abrangente objetivando a verificação quantitativa da metodologia.

Os resultados serão medidos através de comparações entre duas grandezas. A primeira é o número de visitas originadas por consultas feitas em motores de busca. A segunda diz respeito ao número de conjuntos de palavras-chave distintos utilizados em consultas que levaram visitantes aos sítios.

Por exemplo, suponha as consultas +genealogia por dna+ e +dna+; suponha também que o número de visitantes que tenham alcançado o sítio após realizar a primeira consulta tenha sido igual a 10 e aqueles que alcançaram o sítio usando a segunda consulta tenha sido igual a 5. Nesse caso, o número de visitas originadas por consultas em motores de busca é igual a 15 e o número de conjuntos distintos de palavras-chave é igual a 2 (+genealogia por dna+ e +dna+).

Espera-se que o sítio otimizado tenha melhor desempenho nas duas grandezas quando comparado ao seu par não otimizado.

O capítulo se divide em 3 (três) seções. Na Seção 6.1 a metodologia de avaliação é apresentada. Na Seção 6.2 os resultados encontrados são discutidos. Na Seção 6.3 alguns problemas encontrados são apresentados e discutidos.

6.1. Metodologia de avaliação

A avaliação da metodologia de otimização de sítios é feita comparando-se um sítio otimizado com um sítio não otimizado que possuam o mesmo assunto. A preparação de um conjunto de teste é feita em 2 passos:

1. Criação de um sítio com conteúdo relevante a respeito de algum assunto arbitrário. Neste primeiro passo não deve haver qualquer preocupação com a otimização de sítios;

2. Criação de duas versões do sítio obtido como resultado do primeiro passo: uma versão não otimizada, que será simplesmente uma cópia desse sítio, e uma versão otimizada, obtida aplicando-se a metodologia descrita no Capítulo 5 com a adaptação dos planos de Ranganathan.

Após esses dois passos, é necessário publicar os sítios para que ambos possam ser coletados e indexados pelos motores de busca. Conforme discutido na Seção 5.1, a melhor forma de indexar um sítio é através da obtenção de apontadores de páginas que sejam indexadas regularmente pelos motores. Para evitar que uma página seja coletada e indexada enquanto a outra não, escolhe-se uma página adequada aos propósitos do teste e posiciona-se na mesma um apontador para a página principal de cada um dos sítios do conjunto de teste. É importante notar que, dessa forma, a etapa A da otimização de sítios: "fazer com que uma página seja indexada", não é avaliada.

Após a publicação deve-se esperar determinado período de tempo variável para que os sítios sejam indexados e para que as consultas feitas pelos usuários gerem

visitas relacionadas ou não ao seu assunto. É importante ter em mente que uma consulta a um motor de busca não significa necessariamente uma visita ao sítio, um usuário pode realizar diversas consultas a um motor e não necessariamente acessar o sítio, seja porque nenhuma referência a uma página do sítio foi retornada como consulta ou porque o usuário preferiu não acessar a página referenciada no resultado.

Para os experimentos desta dissertação foram criados dois conjuntos de testes: um a respeito do assunto "Genealogia por DNA" e outro a respeito de "Otimização de sítios para motores de busca". Escolheu-se a página principal da empresa Paradigma, sob a qual o pesquisador tinha controle para posicionar os apontadores; essa escolha considerou a freqüência com que tal página é indexada por diversos motores e o alto valor do Pagerank, 5, apresentado pela ferramenta "*Google Toolbar*".

As estatísticas de visitação dos sítios foram coletadas durante o período compreendido pelas datas 28/09/2004 – 04/11/2004, ao final do qual foram analisadas conforme é apresentado a seguir.

Os números apresentados foram gerados pelo programa Awstats (Destailleur). Quando um usuário acessa um sítio, seu navegador envia diversas informações para o servidor, como, por exemplo, o endereço da página cujo apontador referenciou e originou a visita (caso exista tal página). Quando um usuário realiza uma consulta em um motor de busca, o motor cria dinamicamente uma nova página específica para aquela consulta e os termos da consulta fazem parte do endereço dessa página e caso o usuário escolha uma das referências esses

Conjunto 2: Genealogia por DNA

	Sítio Otimizado	Sítio Não Otimizado
Número de conjuntos distintos	105	42
Número de visitas	308	93

Tabela 5 Número de conjuntos de palavras-chave distintos e visitas do Conjunto de teste 2.

Visitas geradas por cada motor de busca:

Motor de Busca	Sítio Otimizado	Sítio Não Otimizado
Google (Google)	282	84
Buscador Terra (Terra)	20	4
Zoom Globo (Globo)	3	0
Go.com (Go)	1	0
Yahoo! (Yahoo)	1	3
Gigabusca (Gigabusca)	1	0
Sapo (Sapo)	0	2

Tabela 6 Motores de busca que geraram visitas para o conjunto de teste 2.

Em ambos os conjuntos de testes o sítio otimizado recebeu um número maior de visitas originadas por motores de busca, além de ter sido alcançado por um número maior de conjuntos de palavras-chave distintos relacionados com seu assunto, o que mostra a eficácia da metodologia proposta.

Porém, houve diferenças entre os resultados obtidos para cada conjunto de teste. No Conjunto 1, um único motor originou visitas (o Google), o número de conjuntos distintos foi 22 e apenas 37 visitas foram devidas a motores. Já no Conjunto 2, 308 visitas foram devidas a 105 consultas distintas em 6 motores. Uma explicação para isso é que as consultas relacionadas com o Conjunto 2 não são tão

disputadas quanto as consultas do Conjunto 1; "consultas disputadas" são consultas que possuem grande número de sítios interessados em aparecer bem para as mesmas – e preparados para tanto – no ranking gerado pelos motores de busca. Outra possível explicação é que o número de consultas relacionadas com o Conjunto 2 é maior que o número de consultas relacionadas com o Conjunto 1.

De fato, o experimento realizado mede a eficácia da metodologia proposta apenas quando compara sítios que possuam o mesmo conteúdo e apontadores externos. São necessárias pesquisas posteriores para comprovar com mais rigor a eficácia da metodologia quando compara sítios que possuam conteúdos e apontadores externos distintos.

6.3. Problemas encontrados

O experimento apresentado foi relativamente restrito, devido sobretudo à dificuldade de se construir um ambiente de testes mais extenso e ao mesmo tempo rigoroso e cientificamente controlado. Seguem abaixo algumas das dificuldades enfrentadas:

- Para que se obtenha testes mais seguros é necessário construir maior número de sítios otimizados e não otimizados com conteúdos relevantes e realistas;
- Os sítios precisam ser indexados, o que exige a obtenção de apontadores, de preferência o mesmo apontador para cada conjunto de teste. No caso de poucos sítios não é tão complicado, porém para um grande número de sítios com duas versões cada, e com conteúdo praticamente idênticos,

pode-se tornar difícil conseguir apontadores relevantes que garantam os testes;

- O tempo entre o sítio ser colocado no ar, indexado e os resultados analisados precisa ser maior. Essa dificuldade se relaciona basicamente com a falta de controle rigoroso do ambiente geral de testes; uma página pode ser coletada e indexada tanto em um curto período de tempo, quanto em um longo período, podendo até não ser nem coletada ou indexada. Tudo irá depender da política de cada motor de busca e da qualidade dos apontadores que a página recebe.

7. Conclusão

Este trabalhou apresentou uma metodologia para a realização da otimização de sítios para motores de busca baseando-se em conceitos da Ciência da Informação, mais especificamente os planos de indexação de Ranganathan. A otimização de sítios é definida como o processo de adequação das páginas de um sítio para que estas apareçam na primeira página de resposta que os motores de busca retornam para consultas de interesse.

Para entender com mais clareza o que é otimização de sítios e como ela pode ser realizada é necessário entender, mesmo que superficialmente, como funciona um motor de busca. Tal explicação foi objeto do Capítulo 3, tendo como base o processo geral de Recuperação de Informação, delineado no Capítulo 2.

O Capítulo 4 definiu detalhadamente o processo de Otimização de Sítios e o diferenciou da prática de spam, com a qual é às vezes confundido. O Capítulo 5 propõe uma metodologia para a realização desse processo. A metodologia divide-se em duas etapas: fazer com o que o sítio seja indexado (etapa A) e fazer com que o sítio apareça para consultas de interesse (etapa B). A segunda etapa baseia-se nos planos de Ranganathan para indexação (cf. Seção 5.2.1) e engloba tanto a seleção de palavras chave quanto o posicionamento das mesmas.

A metodologia proposta nesta pesquisa foi testada e os resultados, assim como a descrição do teste, foram apresentados no Capítulo 6.

Com base nos resultados do experimento e nas discussões apresentadas nesta dissertação, pode-se apresentar como principais conclusões e contribuições deste trabalho:

1. A otimização de sítios é uma tarefa factível. O experimento realizado mostra que os resultados da otimização de sítios não são apenas especulativos, um sítio otimizado pela metodologia desenvolvida nesta pesquisa terá maior visitação gerada por motores de busca do que um que não tenha sido otimizado.

2. A pesquisa mostra que a otimização de sítios pode ser dividida em etapas que empregam profissionais com habilidades diferentes, inerentes à área da Ciência da Informação, porém complementares. Dessa forma é possível a criação de equipes responsáveis pela otimização de sítios, com cada integrante da equipe se especializando em uma tarefa.

3. Apesar da otimização de sítios ser uma área com inúmeros trabalhos comerciais sérios, principalmente no exterior, contribuições acadêmicas são escassas; esta dissertação de certa forma ajuda a preencher essa lacuna e abre caminho para pesquisas mais profundas sobre o tema.

Um problema encontrado no decorrer da pesquisa foi a da criação do experimento, ou seja, o ambiente de testes da metodologia. Para a realização de testes são necessários sítios com conteúdo relevante e que estejam sob o controle total do pesquisador. Por esses motivos o teste realizado foi qualitativo.

Outro problema em relação aos testes diz respeito ao que será comparado. Neste trabalho, foi escolhido demonstrar a eficácia da metodologia comparando dois sítios com conteúdo e apontadores externos inicialmente idênticos, um dos sítios foi otimizado utilizando a metodologia proposta, enquanto o outro permaneceu inalterado. Esse teste demonstra que um sítio pode ser otimizado e com isso obter melhores resultados em motores de busca, porém resta ainda determinar qual é a melhora do sítio otimizado em relação aos demais sítios que também estejam indexados e sobre os quais o pesquisador não possui controle.

Dessa discussão, surgem naturalmente possíveis trabalhos futuros visando aprofundar e expandir esta pesquisa:

1. Realização de testes quantitativos da metodologia, seguindo o esquema de testes usado nesta pesquisa (conjunto de pares de sítios otimizados e não otimizados);

2. Realização de testes qualitativos e quantitativos da eficácia da metodologia quando em comparação com outros sítios que não estejam sob o controle do indexador. Esses testes conseguiriam avaliar todas as etapas da otimização, desde a etapa de obtenção de apontadores até o posicionamento das palavras-chave;

3. Detalhamento maior de cada etapa e sub-etapa da metodologia. Apesar da metodologia descrever os passos necessários para a otimização de sítios em hora nenhuma se pretendeu detalhá-los à exaustão. Isso porque, além da evolução dos motores poder tornar qualquer detalhamento obsoleto, ainda existe o fato de que cada etapa poder ser melhor especificada por

profissionais de áreas distintas e com métodos próprios já consolidados. Por exemplo, a execução do Plano Idéia pode ser feita tanto por profissionais da ciência da informação (análise de assunto), quanto por profissionais de publicidade (determinação do público alvo de um sítio), ou, o que seria mais apropriado, por profissionais de ambas áreas.

4. Criação de ferramentas de auxílio ao otimizador de sítios. Por exemplo, uma ferramenta que realize consultas periodicamente e verifique a posição das páginas de um sítio, alguma ferramenta que sugira palavras-chave baseando-se nas consultas realizadas nos motores e disponibilizadas pelos mesmos; além de outras. Tais ferramentas seriam bastante úteis ao processo, tornando-o o mais automatizado possível e apoiando os profissionais que o executam.

8. Referências Bibliográficas

Anick, Peter. Using terminological feedback for web search refinement: a log-based study. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, p 88-95. 2003.

Amitay, Einat; Carmel, David; Darlow, Adam; Lempel, Ronny; Soffer, Aya. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In: 14th ACM Conference on Hypertext and Hypermedia. Agosto 2003.

AWStats - Free log file analyzer for advanced statistics (GNU GPL). Disponível em: http://awstats.sourceforge.net/

Barroso, Leonardo; Dean, Jeffrey; Hölzle, Urs. WEB Search for a planet: The Google cluster architecture. In: IEEE Micro, Vol. 23, N. 2, pp. 22-28. March/April 2003.

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. Modern Information Retrieval. New York: ACM Press. 1999.

Broder, Andrei. A Taxonomy of Web Search. In: ACM SIGIR Forum. Fall 2002. v.36. n.2.

Buckland, Michael. Information as thing. Journal of the American Society for Information Science. volume 3, number 5, p. 351-360, jun. 1991.

Cho, J; Garcia-Molina, H; Page, L. Efficient crawling through url ordering. In: Proceedings of WWW7, p 161-172, Brisbane. 1998.

Cornelius, Ian. Theorizing Information for Information Science. In: Cronin, Blaise. Annual Review of Information Science and Technology. New Jersey: Information Today, inc. 2002. Volume 36, cap. 9. 393-500.

Davison, Brian D. Recognizing Nepotistic Links on the Web. In: AAAI-2000 Workshop on Artificial Intelligence for Web Search. Julho 2000.

Destailleur, Laurent. AWStats - Free advanced log file analyzer for web, ftp or mail statistics (GNU GPL). Disponível em: http://awstats.sourceforge.net/. Acesso em: 29/11/2004.

Dwork, Cynthia; Kumar, Ravi; Naor, Moni; Sivakumar, D. "Rank aggregation methods for the web". In Proc. 10th International World Wide Web Conference, pp. 613-622. 2001.

Fetterly, Dennis; Manasse, Mark; Najork, Marc. Spam, Damn Spam, and Statistics: Using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS. Junho 2004.

Google Toolbar. Disponível em http://toolbar.google.com/firefox/. Acesso em: 29/11/2004.

Henzinger, Monika R.; Motwani, Rajeev; Silverstein, Craig. Challenges in web search engines. ACM SIGIR Forum, Volume 36 issue 2. Setembro 2002.

Koster, Martijn. Guidelines for Robot Writers. 1993. Disponível em: http://www.robotstxt.org/wc/guidelines.html. Acesso em: 29/11/2004.

Kleinberg, J. Authoritative sources in a hyperlinked environment. In Proceedings of th 9th Annual ACM-SIAM Symposium on discrete algorithms, p 668-677. 1998.

Lancaster, W. F. Vocabulary control for information retrieval. Virginia: Information Resources Press. 2 ed. 1986.

Lancaster, Wilfrid F.; Warner, Amy J. Information retrieval today. Virginia: Information Resources Press. 1993.

Meadow, Charles T. Text information retrieval systems. New York: Academic Press, inc. 1992.

Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Libraries Working Paper, 1998.

Ranganathan, S. R. Elements of library classification. 3.ed. India: Asia Publishing House, 1962. p.88, 89.

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

Salton, Gerard; McGill Michael J. Introduction to Modern Information Retrieval. New York: McGraw-Hill Book Co. 1983.

Silverstein, Craig; Henzinger, Monika; Marais, Hannes; Moricz, Michael. Analysis of a Very Large AltaVista Query Log. Technical Note. Digital Systems Research Center. Outubro 1998.

Svenonius, Elaine. Design of controlled vocabularies. In: Kent, Allen. Encyclopedia of library and information science. Volume 45, supplement 10. p. 82-109. New York: Marcel Dekker, Inc. 1989.