Universidad de Salamanca
Departamento de Informática y Automática

A INCIDÊNCIA DE

WEB SPAM

NOS SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO

Armando Carvalho

DIRECTOR DE TESE

Dr. D. José Luís Alonso Berrocal

VNiVERSiDAD
D SALAMANCA

Universidad de Salamanca

Departamento de Informática y Automática

A incidência de Web Spam nos sistemas de Recuperação de Informação

Tese de Doutoramento

elaborada por

Armando Carlos Costa Carvalho

Director:

Documento revisto por

Dr. D. José Luis Alonso Berrocal

Janeiro 2010

VNiVERSiDAD
D SALAMANCA

UNIVERSIDAD DE SALAMANCA
Departamento de Informática y Automática

A incidência de Web Spam nos sistemas de Recuperação de Informação

Tese de doutoramento apresentada por:

Armando Carlos Costa Carvalho

Dirigida por:

Dr. D. Jose Luis Alonso Berrocal

O doutorando

Salamanca, Janeiro de 2010

Jose Luis Alonso Berrocal, *Profesor Titular de Universidad del Departamento de Informática y Automática de la Universidad de Salamanca*

HACE CONSTAR: *Que* D. Armando Carlos Costa Carvalho, *Licenciado in Informática por lo IPA - Instituto Politécnico Autónomo em Lisboa (Portugal) ha realizado bajo mi dirección la Memoria que lleva por título* La incidencia del Web Spam en los Sistemas de Recuperación de Información, *con el fin de obtener el grado de Doctor por la Universidad de Salamanca.*

Y para que surta los efectos oportunos firmo en Salamanca, a trinta de enero de dos mil diez.

Agradecimentos

Há pessoas que pela sua importância neste projecto e na minha vida foram a mola impulsionadora que garantiram a minha motivação e empenho durante todas as fases do estudo.

Em primeiro lugar quero agradecer à pessoa que mais de perto seguiu este trabalho, ao meu director, o Dr. D. Jose Luis Alonso Berrocal. Um obrigado muito especial pela ajuda, conhecimentos e abertura de novos horizontes e apresentações a outros investigadores desta área de conhecimento.

De uma forma muito especial agradeço à minha querida esposa Graça e aos meus queridos filhos Hugo e André, pela forma como me motivaram e souberam esperar pelos momentos de família, quando dela me afastava para me absorver nos conteúdos da tese. Só o amor que me dedicam pode explicar essa confiança que em mim depositaram.

Agradeço ao meu colega de longas viagens até Salamanca, o José Filipe Lopes, pelo encorajamento que a sua determinada juventude sempre me inspirou.

Agradeço a todos os familiares que, compreendendo as minhas ausências, mantiveram o apoio ao meu projecto e a confiança na capacidade de conseguir chegar ao fim.

Com carinho agradeço ao meu grupo de tertúlia semanal e demais amigos do Centro Cultural de Montefuste.

Bem Hajam.

Aos meus queridos pais - lá no Céu.

A Deus

Resumo

A informação disponível em biliões de páginas na web é um dos factores primordiais para o sucesso crescente do ambiente WWW, mais conhecido como ambiente WEB.

Os motores de pesquisa tornaram-se a 'porta principal' para aceder a essa informação. Por isso, os seus proprietários desenvolvem cada vez mais esforços, preparando especialistas em algoritmos de alta performance para ajudar as pessoas a encontrar as páginas com conteúdos mais relevantes para as suas necessidades.

Conhecer a eficiência dos sistemas de Recuperação de Informação (RI) é importante não só para quem procura por matérias / informação específica em sistemas com imensos conteúdos, mas também para quem tem que garantir um grau de certeza elevado nas respostas dadas às pessoas que, digitalmente, as solicitam.

Especialistas em motores de pesquisas investigam continuamente no sentido de encontrarem algoritmos e construirem sistemas que contribuam para essa mesma ideia - que ajudem as pessoas a encontrar as páginas mais relevantes aos seus interesses particulares [Wu, 2007].

Mais emocionante e determinante se trata quando sabemos que as pessoas (que efectuam as perguntas ao sistema) apenas demonstram verdadeiro interesse pelos primeiros 10/20 links aconselhadas pelos motores de pesquisa [Borodin *et al.*, 2005; Jansen *et al.*, 2000; Silverstein *et al.*, 1999], ainda que as respostas possam ser de milhares.

Daqui surge uma grande competição pela obtenção dos primeiros lugares na lista de respostas que os motores de pesquisas apresentam às pessoas, principalmente pela grande expectativa de gerar lucros que esses lugares podem representar.

Veremos neste trabalho que para alcançar esse objectivo - bom ranking - há várias técnicas que alguns autores classificam de *éticas (white-hat)* ou *menos éticas (gray-hat ou black-hat)*[Wu, 2007], consoante se aproximam mais ou se afastam da pureza e singularidade dos conteúdos.

Estas técnicas de iludir os algoritmos dos motores de pesquisa, conhecida vulgarmente por 'spam' [Gyongyi & Garcia-Molina, 2005; Perkins, 2001] tem vindo a sedimentar o seu conceito na referência a tudo o que de forma não solicitada, tansforma em importante (do ponto de vista de quem pesquisa) um texto ou mensagem.

De entre os diversos métodos de ataque, o 'spam' inventa novas formas, como seja a nova oportunidade de enviar mensagens não solicitadas, de forma maciça, usando o e-mail. Outras variantes de spam fazem-se notar em outras áreas como as mensagens instantâneas (*spim*), por VOIP (*spit*), pelos telemóveis, etc [Becchetti *et al.*, to appear (In Press)].

A web não consegue escapar a esta lista e, na impossibilidade de usarem comunicação directa com o utilizador [Becchetti *et al.*, to appear (In Press); Castilho *et al.*, 2006], os spammers tentam enganar os motores de pesquisa, com técnicas que são conhecidas por 'spamdexing' [Gyongyi & Garcia-Molina, 2005].

Porque os sistemas de RI em documentos de texto guardados digitalmente tratam, essencialmente, de indexação, pesquisa e classificação dos documentos, colocam-nos no ambiente natural do 'spamdexing'.

A nossa análise às técnicas que dificultam estes procedimentos, produzindo ilusórios *rankings*, acrescem imperativos para que se continuem a mostrar como credíveis, de uma forma particular, os primeiros links apresentados nas respostas, uma vez que são esses a que o utilizador mais acede [Marcondes & Sayão, 2002; Saito *et al.*, 2007; Silverstein *et al.*, 1999; Wu, 2007].

Daí que o estudo da evolução de múltiplos algoritmos de ranking, associada à necessária validação dos resultados à perspectiva humana se tenha tornado o enfoque principal da nossa investigação.

As conclusões abrem-nos imensas fronteiras no campo da investigação em que, como sempre, se prevê seja continuada a luta entre os que pretendem a divulgação de todas as tecnologias em favor do ser humano e os outros que se servem dessas oportunidades para construir novas oportunidades de negócio.

É o eterno ciclo.

Abstract

Knowing the efficiency of the Information Retrieval (IR) is important not only for people who look for specific areas in systems with large digital content information, but also for using these systems and has the task of ensuring that the responses are processed optimally to the needs of those who carried out the question.

The specific area that we will address, is located in the search and navigation via the Internet, and the tools are the well-known search engines.

IR systems in text documents stored digitally deal essentially with indexing, search and classification of these documents in response to queries made by users.

A system for IR, current indexing, builds an index that represents the collection of documents.

This index is built by units of indexing (descriptors) for the corresponding weights of certain relevance (normally related to the place where the information appears on pages Web-Title, header, link, etc.).

During the search, we sought descriptors, and assign a value to each relevant document in relation to the consultation, and eventually the documents are ranked by decreasing the values.

It is precisely at the time of calculation of the representativeness of the terms, when checked interference caused by the introduction of foreign elements in the systems, known as a source for the calculation algorithm, which causes changes in the final result, both at indexing, as the inevitable response, unless the user specifies.

The evaluation of RI is therefore a new need, or is crucial to evaluate

the measures for assessing relevance of documents with regard to the need for information expressed in the consultation home. All methods that determine the assertiveness of the search algorithms RI site is called spam.

Our analysis of Web Spam refers to the pages using techniques to fool search engines, giving management end higher, and thus enhancing access to its website. The increase in the final ranking of the pages is of vital importance because, as has been observed in several studies, and evaluates the user accesses mainly to the first links identified by search engines.

To reach this goal, there are several techniques that allow these improvements, and some authors classified as ethical (white-hat) or less ethical (gray-hat or black-hat). The first focuses on improving results improve the content and quality of the pages in the sense of adapting information to the actual needs of users; techniques black-hat use less orthodox methods and the techniques for stuffing, keyword stuffing, cloaking, Web farming, among others as defined Gyomghi and Molina Web Spam Taxonomy.

Search engines, in general, how much they need to improve detection, even because some search engines do not use filtering techniques to exclude any Web spam after it built its indexes.

The study of the multiple ranking algorithms evolution, coupled with the necessary validation of the results to the human mind has become the main focus of our investigation.

The conclusions offer us huge new opportunities to research in which, as always, is expected to continue the struggle between those who want to disclose all the technologies in favor of human beings and others who use these opportunities to build new business ocasions. Is the eternal cycle!

Indice

Indice de Tabelas

Indice de Figuras

Capítulo 1

Introdução

1.1 Objectivos

O estudo científico do conceito de RI já existe há algumas décadas e, segundo Lancaster [Lancaster, 1993], apresenta duas linhas principais de desenvolvimento. A primeira tem as suas origens nos grandes sistemas de bases de dados desenvolvidas pelas instituições americanas: National Library of Medicine (NLM), Department of Defense (DoD) e pela National Aeronautics and Space Administration (NASA), que indexavam as suas bases de dados referenciais utilizando os *thesaurus* específicos das suas áreas temáticas. A segunda linha desenvolveu-se no campo do direito e envolvia a geração da base de dados com o texto completo das leis.

Em concreto, a RI tem como objectivo [Sánchez, 2006] obter, desde uma colecção de documentos, **aqueles que satisfaçam uma necessidade específica do utilizador, de tal forma que a maior parte dos documentos recuperados sejam relevantes para essa necessidade**. É também importante a associação indispensável de RI a Recuperação de Dados (RD), cujos conceitos são abordados por Sánchez [Sánchez, 2006] e também do ponto de vista da relevância da informação.

Sobre a questão da Problemática da RI, refere Bordignon [Bordignon & Tolosa, 2007] que o problema pode ser estudado de dois pontos de vista: **o computacional e o humano**.

Enquanto o primeiro caso - **o computacional** - tem a ver com a construção das estruturas de dados e algoritmos eficientes que melhorem a qualidade das respostas, o segundo - **o humano** - corresponde ao estudo do comportamento e das necessidades dos utilizadores.

Esta necessidade está a evoluir no sentido da interpretação semântica da pergunta associada à pesquisa e que começa a ser uma área de investigação para os algoritmos dos motores de pesquisa de novas gerações.

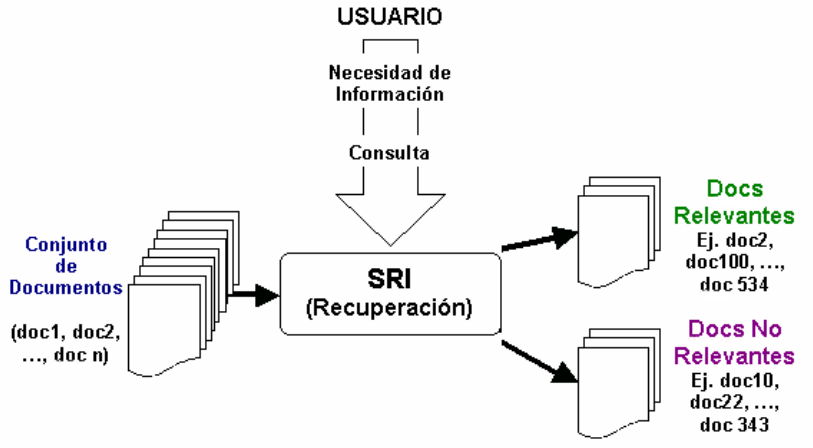

Figura: 1.1: Problematica de RI conforme Bordignon [Bordignon & Tolosa, 2007]

Ao analisarmos a problemática da RI (Fig. 1.1), desde um elevado nível de abstracção, verifica-se que:

- Existe uma colecção de documentos que contém informação realmente de interesse sobre variadíssimos temas
- Existem utilizadores interessados em aceder a essa informação
- Como resposta, o sistema retorna uma lista ordenada de referências a documentos considerados relevantes para o utilizador / pergunta.

Do ponto de vista do utilizador, a resposta ideal deveria conter apenas *links* para páginas que realmente respondessem ao solicitado, mas isto não é uma questão fácil. Entre outros constrangimentos existe um

problema de compatibilização entre o que é pedido e o que está disponível, mesmo ao nível da subjectividade da pergunta.

Apresenta-se então a necessidade de que a resposta seja '**precisa**' de vários pontos de análise, o que pode levar a um conjunto de soluções, em vez de apenas a uma resposta.

Aqui surge então como tendo importância máxima a arquitectura dos sistemas de recuperação de informação (SRI), principalmente no momento da indexação por ser esta a área base e à qual acedem os algoritmos de pesquisa.

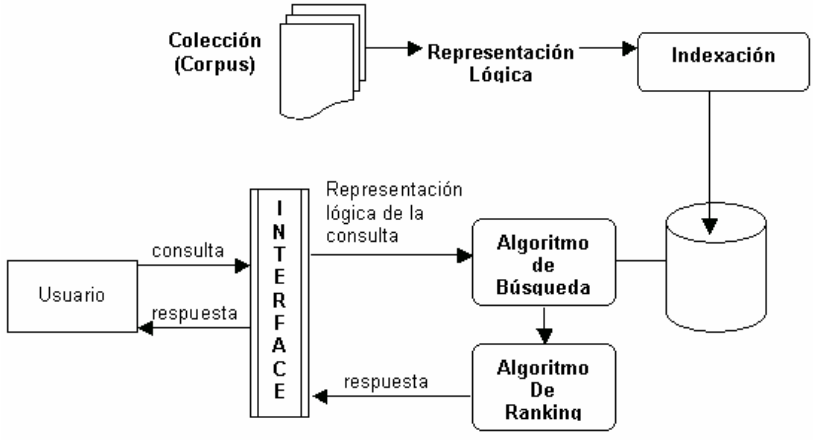

Figura: 1.2: Arquitectura básica de um SRI [Bordignon & Tolosa, 2007]

Como podemos observar na Fig. 1.2, todo o processo de selecção se inicia num conjunto de documentos disponíveis no ambiente em análise (pode não ser a web), e termina com a resposta ao utilizador, normalmente através de browser.

A resposta deve obedecer a um conceito subjacente de relevância em relação ao que é solicitado na pergunta. É por isso que [Sánchez, 2006] nos apresenta diversos modelos de RI, de entre os quais se destacam os que classifica como Clássicos, os Booleanos, os Vectoriais e os Probabilísticos, que podem ser agrupados conforme a Fig. (Fig.1.3), e que são construídos para responder de forma mais objectiva ao critério de relevância.

Foi durante os anos 70 e 80 que se realizaram os primeiros trabalhos

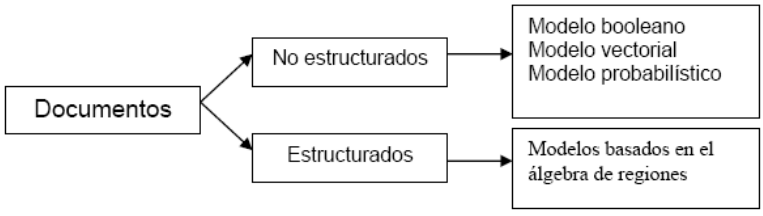

Figura: 1.3: Classificação de modelos de RI [Bordignon & Tolosa, 2007]

[Baeza-Yates & Ribeiro-Neto, 1999; Salton & McGill, 1984; Sánchez, 2006] de introdução de algum deste rigor científico que temos vindo a referir, em complemento de investigações que existiam desde os anos 40 [Sánchez, 2006].

O conceito inicial tem-se vindo a transformar com o tempo, em muito devido ao crescimento exponencial da Web [Kobayashi & Takeda, 2000][1], e verificou-se a todos os níveis: formato dos documentos, tamanho, número de fontes de informação disponíveis, evolução dos recursos de hardware, dos recursos de comunicações, mas, acima de tudo, o alargamento ao comércio electrónico e a novas formas de transacções conducentes à movimentação de capitais.

Esta expansão da Internet, também devida pelas facilidades conseguidas ao nível da conexão dos utilizadores, tornou-se um 'fenómeno' sustentado na evolução tecnológica e que, conforme Schons [Schons, 2007], assenta em dois pressupostos fundamentais:

- **Novas tecnologias** que possibilitaram a interligação de servidores a partir de equipamentos de telecomunicações e computação cada vez mais rápidos;
- **Softwares de comunicação fáceis** de serem utilizados que permitem o acesso e partilha de informação em rede.

Necessariamente que a produção para a Internet passou, para além de ser uma nova realidade, para uma produção preferencial. De um momento para o outro todos querem ter (quase) tudo na Internet, aproveitando as suas características:

[1](A WWW) converteu-se em poucos anos na maior empresa cultural de todos os tempos, equivalente em importância à primeira biblioteca de Alexandria.[Castillo, 2004]

1. **Desregulamentada**: a Internet não possui dono. Não há um manual de regras ou normas de utilização.

2. **Descentralizada e aberta**: Trata-se de uma rede a que todos podem aceder a qualquer hora e praticamente de qualquer lugar. Neste contexto [Martins & da Silva, 2004][2] entendem que, com o surgimento da Internet, o conceito de rede foi significativamente alterado e passou a denotar um sistema aberto capaz de romper fronteiras, permitindo a qualquer indivíduo participar: trata-se de uma estrutura infinita, sem um centro comum e multi-polarizada.

3. **Não hierárquica e interactiva** pelo seu alto grau de interactividade, a Internet promove o remodelamento no estruturado fluxo de informação, possibilitando o desdobramento hierárquico entre emissores e receptores, ao que Lévy [Lévy, 1999] chama de 'inteligência colectiva' e que permite 'reciprocidade na comunicação e a partilha de um contexto'.

Esta nova realidade, principalmente quanto à quantidade de documentos, implica novas ferramentas que permitam, com rigor, responder às necessidades específicas de cada utilizador, com graus de certeza máximos, criando o que Schons considera de 'uma cultura humana de produção' [Schons, 2007].

Fruto de tudo isto passou a haver uma nova necessidade: a de pesquisar em tão vasto universo. Por isso **os motores de pesquisa são considerados a porta de entrada para a web actual** [Wu, 2007].

De uma forma muito rápida, pode-se definir um motor de pesquisa como sendo uma ferramenta que permite às pessoas encontrarem documentos que contenham uma (ou mais vezes) determinada palavra ou frase.

Os motores de pesquisa, conforme referiremos em 1.1.1 são sistemas que exploram na Internet (alguns pesquisam somente na Web, mas outros fazem-no também em News, Gopher, FTP, etc.) quando pedimos informação sobre algum tema, no sentido de fornecerem respostas consentâneas.

As pesquisas são feitas com palavras-chave e/ou com árvores hierárquicas por temas; o resultado da pesquisa é, normalmente, uma lista

[2]Que inclui artigo de Paulo Vaz: Meditação e Tecnologia e referido por Schons

de links que mencionam os temas relacionados com as palavras-chave pesquisadas.

De entre essas ferramentas [Sánchez, 2006] destacam-se os *directórios web* e os *motores de pesquisa* (busca).

As grandes diferenças entre eles devem-se ao facto de que os primeiros[3] disponibilizam a informação organizada de forma temática, enquanto que os segundos[4] são especializados na pesquisa de conteúdos.

Directórios web ou Índices temáticos: São sistemas que permitem pesquisas por temas ou categorias hierarquizadas (embora também incluam sistemas de busca por palavras-chave). Trata-se de bases de dados de links elaborados 'manualmente', isto é, em que há uma prévia intervenção humana na atribuição de cada página à sua categoria/tema.

Motores de pesquisa: São sistemas de busca por palavras chave. São bases de dados que incorporam automaticamente páginas Web mediante 'robôs' de pesquisa pela rede. Estes últimos contêm, geralmente, mais informação que os directórios.

Como iremos ver, infelizmente, estes mecanismos de pesquisa são permissíveis a outras utilizações, como a publicidade, a manipulação da informação e outras formas de controlo do que é mostrado para o utilizador final que efectua a pesquisa.

Enquadrando historicamente a evolução do desenvolvimento dos motores de pesquisa, constatamos que a experiências académicas com o 'Wandex', um web crawler[5] já extinto, ou o 'Aliweb', surgiram os mecanismos de busca desenvolvidos por empresas comerciais. Um dos primeiros terá sido o Lycos (Fig.: 1.4), desenvolvido em 1994.

[3] www.yahoo.es; www.yahoo.com; www.sapo.pt

[4] www.google.com ou .es ou .pt; www.altavista.com

[5] Programa automatizado que acede e percorre os sites seguindo os links presentes nas páginas

Figura: 1.4: Lycos

Depois disso, muitos outros mecanismos de busca apareceram e ganharam adeptos. Entre eles temos o WebCrawler, o Hotbot, Excite, Infoseek, Inktomi, yahoo, Google, AltaVista (Fig.: 1.5), etc. Todos esses mecanismos adicionaram novas tecnologias para aumentar as suas funcionalidades, por vezes de forma muito direccionada para objectivos específicos.

Figura: 1.5: Altavista

O AltaVista, por exemplo, que significa 'uma vista de cima', foi inspirado pela criação de grandes ideias por parte de uma equipe de especialistas com fascínio pela organização da informação. Durante a primavera

de 1995, os cientistas do laboratório de pesquisa da Digital Equipment Corporation situada em Palo Alto, Califórnia, inventaram uma forma de armazenar todas as palavras de todas as páginas HTML da Internet num índice rapidamente pesquisável. Isto levou à criação da primeira base de dados pesquisável, incluindo apenas texto, de toda a World Wide Web.

Outras invenções notáveis da AltaVista incluiram a primeira capacidade de pesquisa multi-linguística presente na Internet e a primeira tecnologia de pesquisa a suportar os idiomas Chinês, Japonês e Coreano[6].

Em 2002, a Yahoo! adquiriu a Inktomi e em 2003 a Overture, a qual era dono da AlltheWeb e AltaVista. Em 2004, a Yahoo! lançou o seu próprio mecanismo de busca baseado na combinação das tecnologias das suas aquisições.

Antes do advento da Web, existiam outros motores de busca para outros protocolos, tais como o motor de busca Archi para sites de FTP anónimos e o motor de busca Verónica para o protocolo Gopher. Hoje, existem centenas de mecanismos de busca. Algumas das novas tecnologias estão presentes em sites como: a9.com, AlltheWeb, Ask Jeeves, Clusty, Gigablast, Ez2Find, Teoma, WiseNut, GoHook, Kartoo, and Vivisimo.

Figura: 1.6: Portal temático e motor de pesquisa

[6]O Babel Fish, foi o primeiro serviço de tradução da Web com capacidade para traduzir palavras, frases ou Web sites inteiros de e para Inglês, Espanhol, Francês, Alemão, Português, Italiano e Russo

Com eles, e aproveitando a evolução tecnológica, respondendo a necessidades crescentes dos utilizadores, tendem a encontrar-se soluções híbridas das duas anteriores: ser simultaneamente um portal temático e um motor de busca (igoogle, yahoo.com), como na figura 1.6.

1.1.1 Conceito de motor de pesquisa

Quando, de uma forma genérica, nos referimos a motor de pesquisa ou de busca, máquina de busca, mecanismo de busca ou pesquisador de facto referimo-nos a um **website especializado em pesquisar e listar páginas da internet a partir de palavras-chave indicadas pelo utilizador**.

A forma desorganizada de inserção de documentos na web, aliada ao tamanho do acervo disponível, torna difícil o acesso aos documentos desejados sem a utilização de sistemas auxiliares. Diante deste cenário, surgiram os motores ou máquinas de pesquisa, e que têm como objectivo central catalogar e organizar a informação contida na web, permitindo que os utilizadores possam pesquisar e recuperar as informações desejadas [Costa & Fernandes, 2007]

Esses sistemas possuem basicamente três módulos:

- **o colector (crawling)** - formado por robôs que são responsáveis pela recolha dos documentos web, em que cada página é um hyperlink para outra página [Kumar *et al.*, 2000; Raghavan *et al.*, 2001], constituindo um grafo de hyperligações que é possível seguir [Caminero & Mikami, 2008].
- **o indexador (indexing)** - que inclui o documento no índice da máquina de origem, disponibilizando-o para consultas e
- **o módulo de consultas (searching)** - que se responsabiliza pela interface com o utilizador, vasculhando o índice da máquina e recuperando o conjunto de documentos que satisfazem a pergunta de busca feita pelo utilizador.

O conjunto de resultados de uma consulta é submetido a **algoritmos de ordenação por relevância**. Estes algoritmos utilizam fórmulas específicas para atribuir valores (métricas) que medem o grau entre a

relevância dos documentos indexados baseados nos termos utilizados na consulta e nas informações contidas em seus índices.

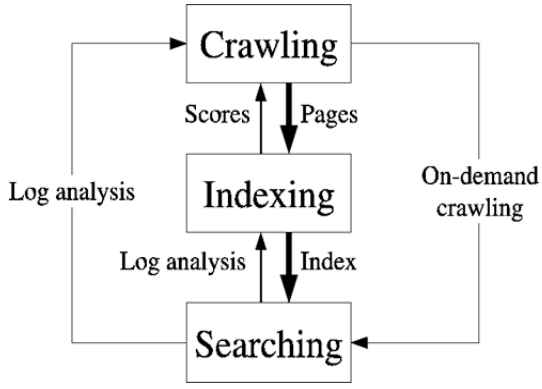

Figura: 1.7: Arquitectura dos motores de pesquisa, mostrando como diferentes componentes podem usar informação gerada por outros componentes. *Fonte: [Castillo, 2004]*

Os motores de pesquisa surgiram logo após o aparecimento da internet, com a intenção de prestar um serviço extremamente importante: a pesquisa de qualquer informação na web, apresentando os resultados de uma forma organizada, e também com a proposta de fazer isto de uma maneira rápida e eficiente. A partir deste preceito básico, diversas empresas se desenvolveram, chegando algumas a valer milhões de dólares. Entre as maiores empresas encontram-se o Google, o Yahoo!, o Lycos e mais recentemente a Amazon.com com o seu mecanismo de busca A9.

Os primeiros motores de busca (como o Yahoo) baseavam-se na indexação de páginas através da sua categorização. Posteriormente surgiram as meta-buscas. A mais recente geração de motores de busca (como o Google) utiliza tecnologias diversas, como a procura por palavras-chave directamente nas páginas e o uso de referências externas espalhadas pela web, permitindo até a tradução directa de páginas (embora de forma tosca) para o idioma do utilizador. O Google, além de fazer a busca pela internet, oferece também o recurso de se efectuar a busca dentro de um site, somente. É essa a ferramenta usada na comunidade Wiki.

Os motores de pesquisa baseiam-se num robô que percorre a Internet à procura de páginas novas para actualizar a sua base de dados automaticamente.

1.1.2 Tipos de motores de pesquisa

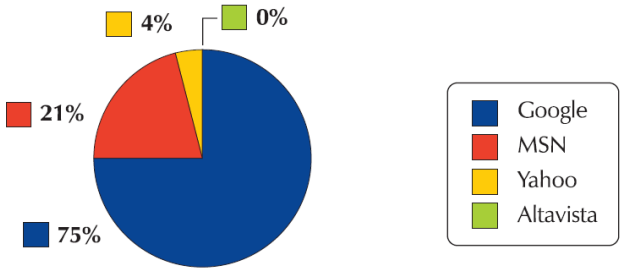

Figura: 1.8: Motores de pesquisa mais utilizadas

Motor	Endereço Web
Google	www.google.com
	www.google.es
Altavista	www.altavista.com
	altavistamagallanes.net
Excite	www.excite.com
	www.excite.es
Hotbot	www.hotbot.com
	hotbot.lycos.com
Infoseek	www.infoseek.com
	infoseek.go.com
Lycos	www.lycos.com
	www.lycos.es
Nothern Light	www.nothernlight.com

Tabela: 1.1: Alguns exemplos de motores de pesquisa.

Existem vários tipos de motores de pesquisa[7]:

- **Globais** são motores que pesquisam a web criando listas organizadas de resposta, consoante o ranking de classificação das páginas

[7]Lista de diversos motores de pesquisa:
$http://byblos.malha.net/component/option,com\_bookmarks/Itemid,47/mode,0/catid,88/navstart,0/search,*/$

encontradas. Os algoritmos de ordenação dessas listas divergem consoante os motores de pesquisa e serão analisados no capítulo 2 deste estudo.

Alguns dos motores de pesquisa mais usados estão referidos na tabela 1.1, e, segundo um estudo de [Costa & Fernandes, 2007], na figura 1.8.

- **Verticais ou temáticos** - Divergem dos motores globais porque enquanto os motores globais devolvem referências a documentos individuais, estes devolvem informação referente à melhor base de dados dentro das quais se realizam pesquisas 'especializadas' de acordo com as suas especializações temáticas. Normalmente este é um serviço pago por mensalidade ou por *clickstream*[8]. Exemplo www.youtube.com

- **Guias locais** são motores com características de abrangência territorial mais reduzida: apenas ao nível local ou regional. Dedicam-se à área comercial e normalmente são um repositório de endereços de empresas e prestadores de serviço do local ou região.

- **Guias de pesquisa local** *ou pesquisador local* têm abrangência nacional e lista as empresas e prestadores de serviços próximas ao endereço do internauta a partir de um texto digitado. A proximidade é avaliada normalmente pelo código postal, ou por coordenadas de GPs. Geralmente os cadastros são gratuitos e apenas os destaques são pagos. É indicado para profissionais e empresas que desejam oferecer os seus produtos ou serviços numa Localidade, rua, bairro ou cidade.

- **Diretórios de websites** são índices de sites, usualmente organizados por categorias e sub-categorias. Tem como finalidade principal permitir ao utilizador encontrar rapidamente sites que desejar, pesquisando por categorias, e não por palavras-chave. Os directórios de sites geralmente possuem uma pesquisa interna, para que os utilizadores possam encontrar sites dentro de seu próprio índice. Os Directórios de Web sites podem ser a nível regional, nacional ou

[8]Clickstream - é o controlo de clicks, em determinadas áreas sensíveis do texto (hiperlinks), que permite a gravação de todos os lugares em que o utilizador clica ao navegar em páginas Web, sendo o resultado gravado no servidor Web ou no próprio cliente. A análise do Clickstream permite aos programadores e analistas várias pesquisas comportamentais dos utilizadores.

Figura: 1.9: Guia local

global, e até mesmo especializados em determinado assunto. Open Directory Project é um exemplo de directórios de sites.

A divulgação de sites de empresas com negócios regionais são acedidos na sua grande maioria quando os profissionais da WEB registam os seus sites nos Motores de Pesquisa Locais para aumentarem as visitas de internautas, em virtude de não haver sistema de actualização automática dos dados que abranja todos os tipos de categorias e com rapidez necessária. Por esta razão, somente cerca de 20% a 25%[9] de tudo que existe na WEB é publicado nos motores de pesquisa.

Para além das soluções híbridas, referidas na introdução, uma novidade [Hoeschl, 2006] desperta com alguma aquidade: os **ontobuscadores**[10], isto é, pesquisadores baseados em Ontologias, como o Ontoweb, entre outras, como veremos mais à frente, em 2.2

A quantidade de informação, a sua dispersão e o grau de mutabilidade tornaram os motores de pesquisa imprescindíveis para o fluxo de acesso e

[9] In *http://pt.wikipedia.org/wiki/Motor_de_pesquisa*

[10] Ontologia ou teoria do conhecimento é um ramo da filosofia que trata dos problemas filosóficos relacionados à crença e ao conhecimento. A epistemologia estuda **a origem, a estrutura, os métodos e a validade do conhecimento** *(in Wikipedia)*

Figura: 1.10: Motor de pesquisa por directorios

a conquista de novos visitantes, e a determinação do que deve ou não ser indexado, como sejam os conteúdos da blogosfera [Kolari *et al.*, 2006b].

Esta habituação dos utilizadores aos motores de pesquisa, [Gyongyi & Garcia-Molina, 2005] na perspectiva de continuarem a prestar uma informação de qualidade, torna necessário que se responda a algumas preocupações.

Entre elas poderemos referir:

- A web continua a crescer muito rápido para que qualquer 'crawler' possa indexar toda a informação.
- Muitas páginas são actualizadas constantemente, o que força o mecanismo de busca a revisitá-la periodicamente.
- Algumas pesquisas são limitadas por palavras-chave, o que pode resultar em falsos positivos.
- Cada vez há mais sites gerados dinamicamente, os quais podem

dificultar ou reduzir a velocidade de indexação, ou podem ainda resultar em excessivas referências para um único sítio.

- Alguns pesquisadores não ordenam as buscas por relevância na pesquisa e sim por comparticipação monetária (taxa).
- Alguns sites utilizam truques para manipular os mecanismos de busca e obterem posições mais favoráveis na classificação final

Principalmente por esta última razão importa que nos debrucemos no capítulo seguinte sobre a temática de Web Spam.

1.2 Conteúdo

O presente trabalho, enquadrando-se dentro das tecnologias de Recuperação de Informação (RI) em ambiente WEB, centra-se no estudo particular das dificuldades causadas por sofisticadas introduções de SPAM [Benczúr *et al.*, 2007a; Gyongyi & Garcia-Molina, 2005] deteorando os resultados das pesquisas efectuadas pelos motores de pesquisa [Castillo *et al.*, 2007a], e encontra-se dividido em 6 capítulos organizados da seguinte forma:

1. Neste primeiro capítulo, além de se descrever a forma global do trabalho, efectua-se uma abordagem concreta ao conceito de Recuperação de Informação de uma forma genérica e em contexto web; também se inicia uma apresentação de alguns conceitos de pesquisa e de motores de pesquisa, indispensáveis à abordagem do tema nos capítulos seguintes.
2. No segundo capítulo apresenta-se o conceito de Spamdexing (ou Web Spam), nomeadamente ao nível de definição e propriedades; apresenta-se uma evolução histórica dos motores de pesquisa e de alguns algoritmos de ranking; evolução de motores de pesquisa e a identificação das suas propriedades fundamentais.
3. No capítulo terceiro referimos alguns conjuntos de estratégias com o objectivo de melhorar a visibilidade dos sites, provocando melhor posição nos resultados naturais de ordenação pelos motores de pesquisa. De entre estes analisaremos a indústria de consultoria dedicada ao SEO - Search Engines Optimization.

4. No capítulo quarto dedicamos especial atenção aos principais tipos de web spam conhecidos e às formas desenvolvidas e em investigação para os combater.

5. No capítulo quinto abordam-se estudos práticos de classificação de sites, no sentido de colaborar com a validação de modelos matemáticos. Apresentamos um caso de estudo relacionado com a forma como os seres humanos podem ser influenciados pelos seus vizinhos (sociedade física ou sociedade digital) na forma como classificam ou evoluem a sua classificação binária e as dificuldades que podem provocar a classificação de 'borderline'.

6. No capítulo seis apresentam-se as principais conclusões do trabalho desenvolvido e as possíveis linhas futuras de investigação que se abrem a partir deste trabalho.

Capítulo 2

Motores de Pesquisa e Algoritmos de classificação

2.1 Introdução

Neste capítulo iniciaremos a abordagem aos conceitos mais directamente relacionados com esta tese, nomeadamente ao nível da identificação dos diversos algoritmos usados durante as fases de *crawling, indexing e searching* dos motores de pesquisa e de como são discutidos os primeiros lugares das listas de classificação, nomeadamente com a introdução de técnicas para iludir os algoritmos.

Importa, no entanto, desde já, retomarmos o conceito de 'motor de pesquisa', a que aludimos no capítulo anterior (Fig. 1.2), que, conforme [Henzinger, 2002], é constituído por três partes distintas:

1. Um 'crawler' que recolhe páginas da web para disponibilizar na colecção de pesquisa do motor (Em termos matemáticos o 'crawler' vê todas a Web como um grafo, em que cada página é um nó e cada link uma haste de conexão);

2. Um 'indexer' que contrói um indice invertido e que é a principal estrutura acedida pelos motores de pesquisa contendo referências a todas as páginas recolhidas pelos 'crawlers';

3. e um 'query handler' que acede ao index e responde ás perguntas

formuladas.

2.1.1 O que é *Web Spam ou Spamdexing*

Comecemos por abordar um conceito estrutural para todo o nosso estudo: SPAM, por ser fundamental no entendimento das razões, que não sendo apenas de escalabilidade, velocidade de processamento e resposta, ou outras de índole técnico, são também razões sociais, que têm motivado um empenho crescente na investigação de novos algoritmos de ranking que permitam manter o grau de confiança entre os utilizadores e os resultados fornecidos pelos pesquisadores às suas consultas [Gyongyi & Garcia-Molina, 2005].

Embora não se possa dizer com total certeza que existe uma única definição para *Web spamming*, também referido por muitos autores como *Spamdexing* [Gyongyi & Garcia-Molina, 2005], é muitas vezes definido como **a prática para conseguir uma posição elevada na lista de classificação dos motores de pesquisa, usando técnicas para enganar os algoritmos de classificação**.

O termo 'spam' conforme referido por Castilho [Castilho *et al.*, 2006], tem sido usado nos últimos anos referindo-se a mensagens *não solicitadas (possivelmente comerciais) e que usam técnicas decepcionantes*, ou como refere [Alliance, 2007] de conteúdo inapropriado, com a possibilidade de conter referências sexuais que tentam enganar os visitantes com falsas promoções e produtos.

Spamdexing é definido por Gyongyi [Gyongyi & Garcia-Molina, 2005] e referido por Castilho [Castilho *et al.*, 2006], como sendo 'qualquer acção com a intenção de conseguir um injustificável aumento de relevância ou importância de uma página Web, considerando o seu real valor'.

Qualquer que seja a definição é certo que *Spam* se refere a **algo indesejável, mesmo perturbador, que influencia negativamente o processo de selecção de informação tratada em ambiente web, com utilização dos protocolos aí disponibilizados**.

Ora, a estrutura de construção do principal protocolo usado na web - o HTTP -, por ser baseado no paradigma pergunta-resposta [Castilho *et al.*, 2006] impossibilita o 'envio' directo de páginas pelos spammers

para os utilizadores finais. Para contornar esta defesa do protocolo os spammers utilizam outras técnicas e meios. Destas a mais utilizada é através de envio de mensagens, aparentemente unidireccionais, via e-mail.

Mas se nos focarmos sobre o modo de operar dos spammers e sobre os processos de RI na web, verificamos que é diferente de todas as outras. Neste caso o alvo principal são os motores de busca e na forma de **enganar e minar** as relações de confiança estabelecidas entre os utilizadores e os motores de pesquisa [Gyongyi & Garcia-Molina, 2005].

Essas técnicas de *Spam* destinando-se aos motores de pesquisa, de facto pretendem obter a atenção dos utilizadores finais, para fins essencialmente comerciais.

Uma das razões, por detrás das dificuldades dos utilizadores em distinguir informações confiáveis de não confiáveis, vem do sucesso que os motores de pesquisa tiveram na última década [Metaxas & DeStefano, 2005]. Os utilizadores têm vindo a aumentar a sua confiança nos pesquisadores como um meio de obter informações, bem como os spamers têm, com êxito, conseguido conduzir essa confiança para os resultados de cada pesquisa.

Para que seja possível continuar a existir confiança nos seus resultados, os construtores dos motores de pesquisa, desenvolveram grande esforço no sentido de proporcionar respostas isentas de *spam*. Nessa perspectiva desenvolveram sofisticadas estratégias de *ranking*. Duas dessas estratégias mais conhecidas e que têm recebido maior atenção, quer ao nível do desenvolvimento quer de melhoramento, são os algoritmos de **PageRank (PR)** [Brin & Page, 1998; Metaxas & DeStefano, 2005] e **HITS** [Kleinberg, 1999; Metaxas & DeStefano, 2005].

Mas, se algumas estimativas indicam que, pelo menos 8% de todas as páginas indexadas são spam [Fetterly *et al.*, 2004; Metaxas & DeStefano, 2005], os investigadores classificam o web spam como o maior desafio actual das pesquisas em ambiente web [Henzinger *et al.*, 2002; Metaxas & DeStefano, 2005], isto porque os motores de pesquisa vêem o web spam como uma interferência às suas operações. Os estudos para limitar esses constrangimentos encontram dificuldades na identificação automática de spam com base apenas em algoritmos matemáticos (graph isomorphism) [Bharat *et al.*, 2001; Metaxas & DeStefano, 2005] . Com efeito precisamos de compreender socialmente a questão do web spam e só depois analisar

as questões técnicas, dado que é na área dos comportamentos sociais que o spam está sempre mais actualizado.

Sabemos como a web mudou o paradigma da informação. Qualquer organização tem um site na web e as preocupações relacionadas com HCI (Human Computer Interaction) [1] são cada vez mais evidentes, no sentido de aumentar o conforto e facilidade de navegação aos utilizadores, bem como criando respostas para as solicitações que possam ter.

A maioria das pessoas, com acesso on-line, utiliza um motor de pesquisa para obter informações e tomar decisões que podem ser dos mais diversos temas: médico, financeiro, cultural, político, de segurança ou outras implicações importantes [Hindman *et al.*, 2003; Lynch, 2001; Metaxas & DeStefano, 2005]. Além disso, **85% das vezes, as pessoas apenas consultam os primeiros 10 resultados da resposta** [Metaxas & DeStefano, 2005; Silverstein *et al.*, 1999]. Perante isto, não é de estranhar que qualquer pessoa com uma presença na web lute por um lugar no top das dez posições mais relevantes dos resultados da pesquisa.

Daí que o objectivo de alcançar um elevado PR se tenha tornado obsessivo para os departamentos de TI de algumas empresas, bem como a principal razão de ser de empresas produtoras de spam.

A importância da colocação no 'top-10' deu origem a um novo sector, que pretende vender know-how para um posicionamento relevante nos resultados de pesquisa e inclui empresas, publicações e mesmo conferências.

Alguns deles estão dispostos a contornar a verdade, a fim de enganar os mecanismos de pesquisa e os seus utilizadores, através da criação de páginas da Web que contenham spam web [Fetterly *et al.*, 2004; Metaxas & DeStefano, 2005].

Os criadores de spam na web podem ser empresas muito específicas que comercializam os seus conhecimentos como um serviço, mas também podem ser os web masters das empresas e organizações que, em qualquer caso, pretendam modificar favoravelmente o resultado do PR. O principal modo de operar dos Spammers é atacando os motores de busca através de texto e [Gyongyi & Garcia-Molina, 2005; Henzinger *et al.*, 2002; Metaxas

[1] Human Computer Interaction (**HCI**), ou Man Machine Interaction (**MMI**) ou Computer Human Interaction (**CHI**) refere-se ao estudo da interacção entre pessoas (utilizadores) e computadores

& DeStefano, 2005] manipulação de links.

Vejamos alguns casos, a que voltaremos com mais detalhe no capítulo 4:

- **Text spam:** Usa técnicas de repetição excessiva de texto e/ou adicionamento de texto irrelevante - em relação ao contexto original da página - que irá causar cálculo incorreto do PR; acrescentando meta-palavras-chave enganadoras ou irrelevantes 'anchor text' que causarão incorrectas aplicações das classificações heurísticas.

- **Link spam:** Esta técnica tem como objectivo alterar a estrutura do webgraph, a fim de causar cálculo incorreto da reputação da página. Esses exemplos são os chamados 'link-farms','mutual admiration societies', 'page awards', 'domain flooding' (multiplicidade de domínios que redirecionam para um site), etc.

Ambos os tipos de spam têm por objectivo impulsionar a classificação das suas páginas web, chegando mesmo a ser utilizadas técnicas de disfarce do próprio spam [Gyongy *et al.*, 2004; Henzinger *et al.*, 2002; Metaxas & DeStefano, 2005], como teremos oportunidade de ver.

Estas páginas podem ser criados estática ou dinamicamente. Enquanto as páginas estáticas, por exemplo, podem usar links ocultos e/ou texto oculto com cores iguais para tinta e fundo ou tamanho de fonte muito pequeno apenas perceptível por um indexador, mas não pelo homem, as páginas dinâmicas podem mudar o seu conteúdo no momento, dependendo do visitante.

Toda esta interactividade, também devida ao facto de que qualquer pessoa poder criar páginas web, originou um problema de credibilidade da informação.

Um público habituado a confiar na palavra escrita de jornais e livros é incapaz de pensar criticamente sobre as informações obtidas a partir da web. Um estudo referido por [Graham & Metaxas, 2003; Metaxas & DeStefano, 2005] apurou que os estudantes universitários que usam a web como uma fonte primária de informação, para além de não referirem mais do que uma única fonte, têem dificuldade em reconhecer fontes fidedignas online. Em especial, dois em cada três estudantes, são incapazes de distinguir entre factos e publicidade.

Muito poucos efectuaram validações externas para verificar a validade da informação obtida. Ao mesmo tempo, eles têm grande confiança nas suas capacidades para distinguir sites fidedignos de não-confiáveis, especialmente quando se sentem tecnicamente competentes. Não temos motivos para acreditar que o público em geral tenha melhor desempenho do que qualquer estudante com formação mais elevada [Corey, 2001; Metaxas & DeStefano, 2005].

2.2 Evolução dos motores de busca e dos modelos de Classificação

2.2.1 Classificação por palavras raras: *tf.idf ranking*

No início dos anos 90, quando na web havia poucos milhões de servidores, a primeira geração de motores de busca classificavam o ranking de resultados da pesquisa usando técnicas clássicas de RI: **quanto mais palavras raras (pouco usadas) dois documentos partilhassem, mais semelhantes eram consideradas** [Henzinger, 2001; Metaxas & DeStefano, 2005]. Deste modo uma pesquisa Q é tratada como um documento curto, no mínimo de uma só palavra, e os resultados de uma pesquisa de Q são classificados de acordo com a sua similaridade com a consulta (palavras raras coincidentes).

$$TF(d,t) = \begin{cases} 0 & \text{if n(d,t)} = 0; \\ 1 + \log(1 + \log(n(d,t))) & \text{otherwise.} \end{cases} \tag{2.1}$$

$$\mathbf{IDF(t)} = \log(1 + |D||D_t|) \tag{2.2}$$

O primeiro ataque a este 'tf.idf ranking' (Formula 2.1 e 2.2), como é conhecido, veio de dentro dos próprios motores de busca. Por volta de 1995, motores de busca começaram a vender palavras-chave para anunciantes como uma forma de gerar receitas: Se uma consulta de pesquisa continha uma palavra-chave 'vendida', os resultados incluíam o anúncio e uma classificação mais elevada para o link do patrocinador do site. Esta é a primeira vez que temos um ranking socialmente inspirado, que segue as práticas de marketing do mundo real.

Misturando os resultados da pesquisa com publicidade paga levantaram-se sérias questões éticas, mas também se mostrou como sendo uma forma de obter lucros para os spammers, que iniciaram os seus próprios ataques, criando páginas com muitas palavras-chave raras para obter um maior ranking.

Com esta técnica os spammers conseguiram confundir a primeira geração dos motores de busca [Metaxas & DeStefano, 2005]: O 'maldoso' associa assim uma ou mais palavras, sem sinais sugestivos de ter alterado os valores da pessoa ou idéias em referência no site.

Para evitar os spammers, os motores de busca tentaram manter secretos os seus algoritmos de ranking. No entanto este segredo não durou muito, pois não conseguiu resistir às novas técnicas de re-engenharia (reverse engineering) [Marchiori, 1997; Metaxas & DeStefano, 2005; Pringle *et al.*, 1998].

2.2.2 Princípio de votos por link

A segunda geração de motores de pesquisa - conforme modelo de Metaxas [Metaxas & DeStefano, 2005] - iniciou uma técnica mais sofisticadas para classificação num esforço para anular os efeitos das 'palavras raras'.

Uma das mais bem sucedidas técnicas baseou-se no princípio da votação por 'link': Cada web site s tem valor igual à sua popularidade, que é influenciada por um conjunto de sites Bs que apontam para esse site s.

O site da Lycos tornou-se o campeão do ranking desta técnica e atingiu o seu próprio pico de popularidade, conforme o mesmo autor [Metaxas & DeStefano, 2005], no ano de 1996. De qualquer forma foi uma técnica positiva dado que conseguiu anular as questões éticas introduzidas pela combinação de publicidade com ranking.

Mas, infelizmente, este método de classificação também não conseguiu parar os spammers. Os Spammers criaram 'clusters' de Web Sites interligados, que tinham conteúdo idêntico ou similar com o site que estavam a promover. Esta técnica veio a ser conhecido como 'link farms' (LF). O princípio do 'voto por link' foi inspirado socialmente, de modo que os spammers utilizaram o método conhecido por 'bandwagon' para

o contornar [Metaxas & DeStefano, 2005]. Com ele, o 'maldoso' tenta convencer-nos de que todos os membros do nosso grupo aceitam o nosso programa e que, por isso, temos de seguir o nosso grupo e 'saltar sobre a carruagem'.

Do mesmo modo, o spammer promove a ilusão de um elevado grau de popularidade criando inter-ligações, controladas internamente, entre muitos sites que acabarão por proporcionar a todos eles alta classificação.

2.2.3 Page Rank (Ranking)

A introdução do método designado por PageRank, em 1998, foi um dos principais desenvolvimentos para os motores de pesquisa, porque ele incorpora maior sofisticação para fornecer uma solução anti-spam. **No PageRank, nem todos os link contribuem igualmente para a reputação de uma página**. Em vez disso, as ligações de alta reputação contribuem de forma muito superior a links de outros sites menos reputados. Dessa forma, as redes de sites desenvolvidas por spammers (clusters e outros) não iriam influenciar muito a sua própria PageRank.

O conceito inerente ao PageRank [Gyongy *et al.*, 2004] é de que **uma página web é realmente importante se várias outras páginas importantes apontam para ela**.

Correspondentemente, PageRank é baseada num reforço mútuo entre as páginas: a importância de uma determinada página *influencia* e é *influenciada* pela importância de outras páginas.

O nível de PageRank $\mathbf{r}(p)$ de uma página p é definido por:

$$\mathbf{r}\,(p) = \alpha . \sum_{q:(q,p)\in\epsilon} \frac{\mathbf{r}(q)}{\omega(q)} + (1-\alpha).\frac{1}{N} \tag{2.3}$$

em que α é um 'decay factor'[2].

A equação matriz equivalente é:

$$\mathbf{r} = \alpha.T.\mathbf{r} + (1-\alpha).\frac{1}{N}.\mathbf{1}N \tag{2.4}$$

[2]Factor de queda

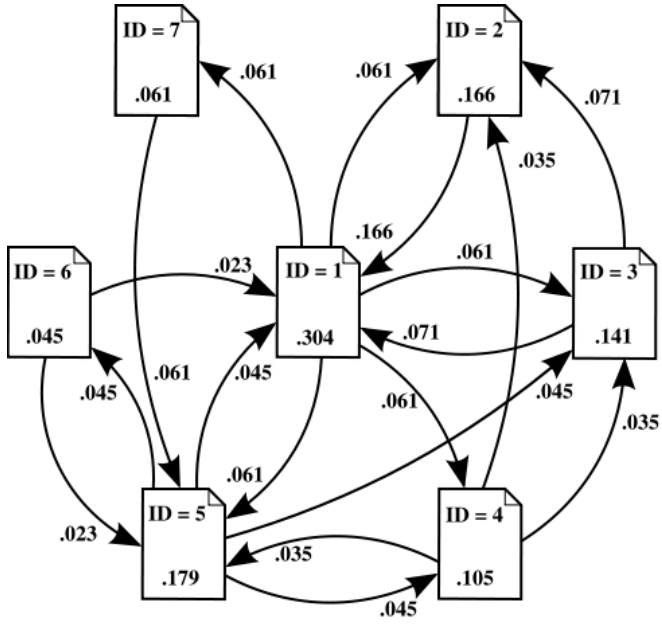

Figura: 2.1: Page Rank em funcionamento

Daí que o resultado de uma página p seja a soma de dois componentes: uma parte da pontuação vem de páginas que apontam para p, e a restante parte (estática) da pontuação é igual para todas as páginas da web.

A pontuação do PageRank pode ser computada de forma iterativa, por exemplo, através da aplicação do método de Jacobi, no entanto, num sentido estrito da matemática, as iterações devem ser executadas por convergência. É mais comum, na prática, a utilização apenas de um número fixo de iterações N.

É importante notar que, enquanto a aplicação do algoritmo base de PageRank atribui a mesma pontuação estática para cada página, esta variante pode quebrar essa regra. Na matriz equação

$$\mathbf{r} = \alpha.T.\mathbf{r} + (1 - \alpha).\mathbf{d} \tag{2.5}$$

o vector d é um vetor de distribuição de pontuação estática arbitrário, não negativo, cujas entradas podem totalizar até um. O vector d pode

ser usado para atribuir uma pontuação estática diferente de zero a um conjunto de páginas especiais, de qualidade assumida. O resultado dessas páginas especiais é, então, espalhado durante as iterações para as páginas que elas apontam.

Os melhores links que um determinado site pode receber são os de páginas de assuntos relacionados ao site, que tenham um alto PageRank e que não forneçam links para outros sites.

A combinação da optimização do site com a popularidade/PR pode resultar numa significativa melhoria no posicionamento do site nos resultados das pesquisas dos principais mecanismos especializados da internet, e isto pode resultar em ficar visível nas primeiras posições com muito mais visitas e consequentemente muito mais possibilidades de novos negócios.

Surge assim o conceito de **popularidade do site** como sendo uma medida da quantidade e da qualidade dos links apontados para ele. Um site que recebe muitos links com origem em outros de conteúdo de qualidade e que também tenham alta popularidade, será considerado de alta popularidade e poderá ser considerado relevante para determinadas pesquisas.

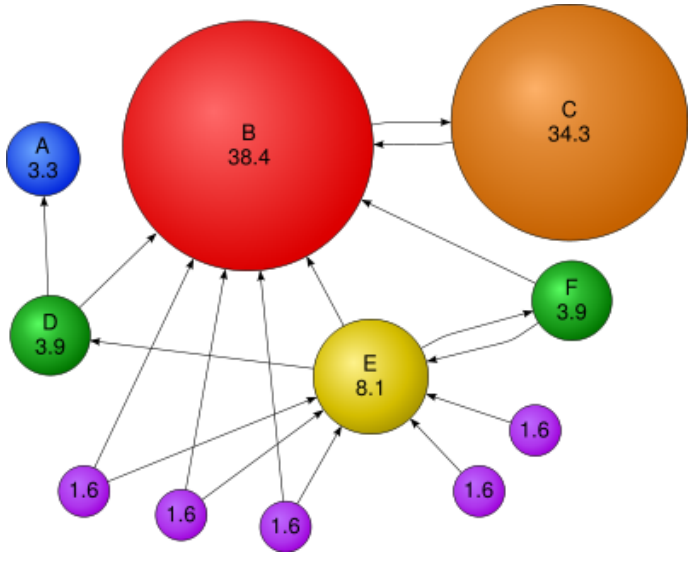

Figura: 2.2: Page Rank [Adaptado de Wikipédia]

Em outras palavras, **um site é considerado importante quando vários outros sites importantes o recomendam através de links**. Cada link é considerado um voto e os votos de sites importantes têm peso maior. Também podemos dizer que os links de sites de conteúdo relacionado ao seu têm peso maior do que sites de conteúdo não relacionado.

2.2.4 O Algoritmo Hilltop

Uma das deficiências do algoritmo de PageRank é que qualquer link, em qualquer página contida no índice, aumenta o PageRank (e melhora o ranking) da página que recebe o link. Entre vários outros, dois problemas maiores preocupavam os investigadores da Google, enquanto utilizadors deste algoritmo.

- Os webmasters começaram a comprar links, que apontassem para eles, para aumentar o próprio PageRank;
- Uma vez construído um site com um valor alto de Pagerank, ficava fácil para os webmasters construírem outros sites e, de imediato, apontar links das suas próprias páginas e conseguir uma boa posição no ranking.

O algoritmo Hilltop [Bharat & Mihaila, 2000][3] responde positivamente à resolução destes dois problemas.

Não há informações oficiais de que o Hilltop tenha sido implementado, no entanto os indicadores dos resultados do Google refletem muitos dos conceitos do algoritmo, o que pode indiciar que o Hilltop (ou boa parte dele) tenha sido incorporado no algoritmo do Google [Gupta, 2003].

A explicação dos seus construtores [Bharat & Mihaila, 2000], é esclarecedora sobre o modo de operar do algoritmo:

> 'A nossa técnica é fundamentada nas mesmas suposições dos outros algoritmos baseados em conectividade, ou seja, que o número e qualidade das fontes que fazem referência a uma página dão uma boa medida da qualidade da mesma.

[3] O Hilltop foi concebido por Krishna Bharat e George Mihaila

> A diferença chave consiste no facto de que nós **consideramos apenas fontes que sejam 'experts'** - páginas que tenham sido criadas com o propósito específico de encaminhar as pessoas aos recursos que procuram. Para responder a uma pesquisa, primeiro computamos uma lista dos maiores experts naquele tópico. A seguir, identificamos links relevantes dentro desse conjunto de experts, seguimo-los para identificar páginas alvo. Essas páginas alvo são então rankeadas, de acordo com o número e a relevância de experts não-afiliados que apontam para elas. Dessa forma, o posicionamento de uma página reflete a opinião colectiva dos melhores experts independentes naquele tópico. Quando um conjunto de experts não existir, o algoritmo Hilltop não retornará resultados. Assim, **Hilltop foca-se na relevância dos resultados, e não na abrangência da pesquisa**.'

O Hilltop procura detectar hosts afiliados, mas se um link apontar para uma página num host afiliado, o valor do link é descontado.

Dois hosts são considerados afiliados, conforme referido por Bharat [Bharat & Mihaila, 2000](secção 2.1), se:

1. Eles têm os mesmos primeiros três octetos de endereço IP, ou seja, os IPs pertencem a uma mesma classe C. Por exemplo, os hosts com IPs 200.109.112.132 e 200.109.112.133 (ou qualquer outro host de IP 200.109.112.xxx) são considerados afiliados.

   Isto foi implementado porque webmasters que possuíam muitos domínios costumavam armazená-lo ou num mesmo IP (é possível ter vários domínios em um mesmo IP), ou numa mesma empresa de webhosting, na qual os grupos de IP são distribuídos por uns poucos grupos de classe C. Isso diminuiu muito o poder dos webmasters que possuíam networks privadas.

2. O primeiro nome não-genérico do domínio é o mesmo. Por exemplo, ibm.com e ibm.co.uk são afiliados (com, co e uk são nomes genéricos; o primeiro nome não-genérico, ibm, é o mesmo). Assim, links em ibm.co.uk ou em research.ibm.com não contribuem para melhorar o ranking de ibm.com. A intenção, igualmente, é impedir que sites promovam outros domínios próximos, por interesses particulares.

3. A relação de filiação é transitiva: se o site A é filiado ao site B, e o site B é filiado ao site C, então o site A é filiado ao site C.

Este melhoramento permite que os links deixem de ter o mesmo peso, pois enquanto que o PR determina a 'authority' de uma página no conceito geral, o algoritmo Hilltop (LocalScore) determina a 'authority' de uma página no escopo específico dos termos de pesquisa.

O artigo de Bharat [Bharat & Mihaila, 2000] define expert (secção 2.2) simplesmente como qualquer página que contenha um número mínimo de links (no exemplo define 5 links, mas na adaptação ao modelo do Google o mínimo é de dois [Gupta, 2003]) para hosts não afiliados.

A Google parece confiar no facto de que os verdadeiros experts são fontes confiáveis de informações (links). Experts em geral são cautelosos na inclusão de links, e é pouco provável que se inclinem ao comércio de PageRank.

A Secção 3, do mesmo artigo [Bharat & Mihaila, 2000], detalha como são calculados os pesos individuais, ficando claro que o resultado é tanto maior quanto maior for o número de experts que apontem para determinada página, e quanto maior for a coincidência de palavras-chave presentes em pontos chave do expert (título, âncora) etc. e na expressão de pesquisa.

O aproveitamento por firmas como a Google deste novo algoritmo permitiu a melhoria dos seus algoritmos particulares.

Veja-se o caso das fórmulas de Ranking do Google antes e depois do Hilltop [Gupta, 2003]:

Antes:

$$(1-d) + a(RS) * (1-e) + b(PR * fb) \tag{2.6}$$

Depois:

$$(1-d) + a(RS) * (1-e) + b(PR * fb) \mathbf{* \ (1\text{-}f) + c \ (LS)} \tag{2.7}$$

Em que:

RS = RelevanceScore: (Pontuação baseadas em palavras-chave que aparecem no título, nas metatags, no texto, URLs, etc)

PR = PageRank: (Pontuação baseada no número e valor dos PR das páginas com links para o site. Fórmula após algumas iterações:

$$\mathbf{PR\ (A)} = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (2.8)$$

ou

$$PR(A) = (1-d) + d\sum_{i=1}^{n} \frac{PR(Ti)}{C(Ti)} \quad (2.9)$$

onde PR de A é a soma dos PR de cada página com links para ele (A) dividido pelo número de links em cada uma dessas páginas. 'd' é um factor (decay factor) igual a 0,15.

LS = LocalScore: (Pontuação computada a partir de documentos 'expert'. Tem variáveis e valores diferentes para o termo de pesquisa que apareçam no título (16), cabeçalho (6), texto âncora (1), termo de pesquisa densidade etc.)[4]

a, b, c = Tweak Weight Controls: ('fine-tuning' regularização de resultados)

d, e, f = Dampener Controls: ('fine-tuning's')

Kumar [Kumar *et al.*, 2000], apresenta algumas melhorias para o algoritmo no conceito de análise da Web como um grafo (graph), indicando as alterações muito específicas introduzidas pelo próprio Bharat [Bharat & Henzinger, 1998] e ainda por Chakrabarti [Chakrabarti *et al.*, 1998a;b] ao nível do conceito booleano das fontes.

2.2.5 TrustRank

O Algoritmo TrustRank, apresentado por Gyongy [Gyongy *et al.*, 2004], pretende separar, de forma semi-automática, os websites creden-

[4]Números entre parênteses são os valores originais, que podem ter sido alterados pelo Google

ciados dos que contenham spam. Toda a sua estrutura baseia-se no conceito de 'Atenuação por Confiança' (*trust attenuation*).

O Trust Rank (incluído nos algoritmos de PR dos principais motores de pesquisa Google e Yahoo) tornou-se um dos factores mais importantes na determinação do ranking real (PR) de um website, também pela inclusão, de forma escondida, de alguns factores que podem influir no Trust Rank de um website[5]. A figura 2.3 apresenta alguns dos Factores que aumentam o TrustRank[6].

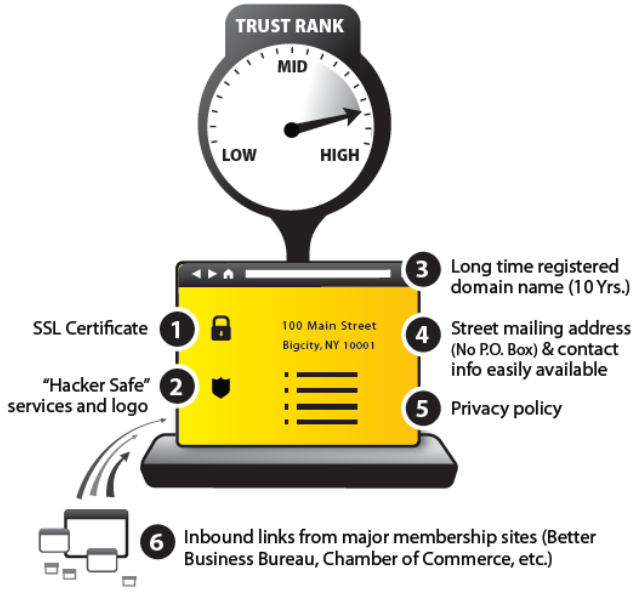

Figura: 2.3: Factores que aumentam o TrustRank

1. **Idade / História do domínio** - A idade do domínio torna-se um factor importante na avaliação de 'confiança', com posibilidade de influenciar o PR global, bem assim como a história do domínio para detectar qualquer atividade SPAM. Isto porque, para melhoria do PR, poderemos interferir em muitas áreas, nomeadamente com técnicas de SEO, mas não podemos alterar a Idade e a

[5]Adaptado de http://blog.seohawk.com/page-rank-vs-trust-rank/

[6]Fonte: http://searchengineoptimization.elliance.com/pdfs/TrustRank-Factors.pdf

História (log) de um domínio.

2. **A quantidade dos 'back-links'** - De forma diferente do PR, neste algoritmo a quantidade das *ligações de volta* não significam necessariamente um aumento de 'confiança'. Isto porque os mecanismos de pesquisa podem facilmente detectar qualquer padrão anormal 'escondido' nessas ligações'. Se, por exemplo, um webmaster comprar 50 'back links' com PR 3-4, qualquer crawler identificará que a confiança não é real.
3. **Conteúdo do site** - A existência de artigos não originais pode provocar uma forte quebra de posicionamento de ranking tanto no Google como no Yahoo. A razão por detrás deste declínio é a duplicação de artigos que são submetidos a várias centenas de websites por vários webmasters e SEO.

Gyongy et al. [Gyongy *et al.*, 2004] identificam que no início está a criação de sementes (seed) que necessitam de, por intervenção de especialistas humanos, serem catalogadas como páginas boas ou páginas contendo spam.

Figura: 2.4: Estrutra de Links

Depois de identificar manualmente a reputação das sementes, utiliza a estrutura de links da web (fig 2.4) para descobrir outras páginas que sejam susceptíveis de ser boas.

No documento em análise [Gyongy *et al.*, 2004], é explicado o algoritmo (fig 2.7) e as suas etapas de implementação, sendo de notar, na análise de pormenor, que a confiança vai reduzindo na medida em que nos afastamos das sementes boas.

De entre diversas formas de conseguir a medição dessa atenuação de confiança, são-nos apresentadas duas possibilidades: Trust Dampening e Trust Splitting.

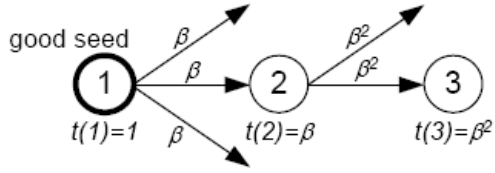

Figura: 2.5: Trust Dampening

A figura 2.5 ilustra a primeira ideia, chamada de amortecedor de confiança '*trust dampening*'. Uma vez que a página 2 está a um link de distância da 'semente boa', a página 1, é-lhe atribuído um amortecedor de confiança de β, em que $\beta < 1$. Quanto à página 3 é acedível à distância de 1 link desde a página 2 atribuímos-lhe um amortecedor de confiança igualmente de β, ou seja o amortecedor de confiança entre 1 e 3 é de $\beta.\beta$.

Igualmente precisamos de decidir como atribuímos idênticos níveis para páginas que contenham muitos links. Por exemplo, na Figura 2.5, é assumido que a página 1 também aponta para a página 3. Poderemos então atribuir à página 3 o máximo grau de confiança, neste caso β, ou, por outro lado, a média dos graus de confiança, que neste caso são $\beta.\beta/2$.

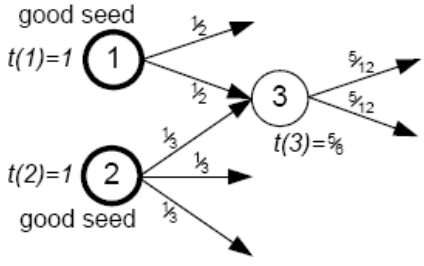

Figura: 2.6: Trust Splitting

Outra técnica possível de utilização, chamada de *trust splitting*, é baseada na seguinte observação: O cuidado com que as pessoas aumentam links para as suas páginas é, frequentemente, inversamente proporcional ao número de links existentes na página. Quer dizer, se uma página 'boa' tem apenas uma mão cheia de links para o exterior, provavelmente as páginas apontadas também serão 'boas'. Contudo, se uma página boa tem centenas de links para o exterior, é muito provável que alguns apontem

para páginas com spam.

Esta hipótese conduz-nos a dividir a confiança na medida em que há propagação de links para outras páginas: se a página p tem um grau de confiança $T(p)$ e ela aponta para $\omega(p)$ páginas, então cada uma das $\omega(p)$ páginas receberá uma fracção de $T(p)/\omega(p)$ de p. Neste caso, o cálculo do valor da página será a soma das diversas fracções recebidas pelos links de entrada.

Intuitivamente poder-se-á dizer que quanto maiores 'créditos' uma página acumular desde outras páginas, maior será a probabilidade de ser uma página boa.

A figura 2.6 ilustra esta segunda hipótese de 'trust splitting'. A página 1, que está considerada como semente boa, tem 2 outlinks, por isso distribui metade do seu 'valor' de 1 por ambas as páginas para onde aponta. Da mesma forma, a boa página 2 tem 3 outlinks, e assim cada página apontada recebe 1/3 do valor de 2. O valor de 3 será então $1/2 + 1/3 = 5/6$.

Veja-se que poderemos ainda combinar os dois métodos (splitting com dampening). Na figura 2.6, por exemplo, a página 3 pode receber um valor de $\beta.(1/2 + 1/3)$.

A implementação do algoritmo é referida na figura 2.7.

2.2.6 Hits

HITS é outra forma socialmente inspirada para determinar o ranking, que também tem recebido muita atenção [Kleinberg, 1999; Metaxas & DeStefano, 2005]. O algoritmo HITS divide os sites relacionados a uma consulta entre 'hubs' e 'autoridades'. Hubs são sites que contêm apontadores para muitas autoridades, ao passo que as autoridades são locais que estão apontados pelos hubs. A referência mútua faz com que ambos ganhem créditos na posição classificativa.

O algoritmo da Teoma[7] é baseado num trabalho chamado HITS, desenvolvido por Kleinberg; o paper original era intitulado Authoritative Sources in a Hyperlinked Environment.

O algoritmo procura classificar cada página em 'hub' (que contém

[7] Teoma significa 'expert' em gaélico

```
function TrustRank
input
        T       transition matrix
        N       number of pages
        L       limit of oracle invocations
        α_B     decay factor for biased PageRank
        M_B     number of biased PageRank iterations
output
        t*      TrustRank scores
begin
        // evaluate seed-desirability of pages
(1)     s = SelectSeed(...)
        // generate corresponding ordering
(2)     σ = Rank({1,...,N}, s)
        // select good seeds
(3)     d = 0_N
        for i = 1 to L do
                if O(σ(i)) == 1 then
                        d(σ(i)) = 1
        // normalize static score distribution vector
(4)     d = d/|d|
        // compute TrustRank scores
(5)     t* = d
        for i = 1 to M_B do
                t* = α_B · T · t* + (1 − α_B) · d
        return t*
end
```

Figura: 2.7: Função para implementação do algoritmo de Trust Rank[Gyongy *et al.*, 2004].

muitos links apontando para outras páginas) ou 'authority' (que recebe muitos links de outras páginas).

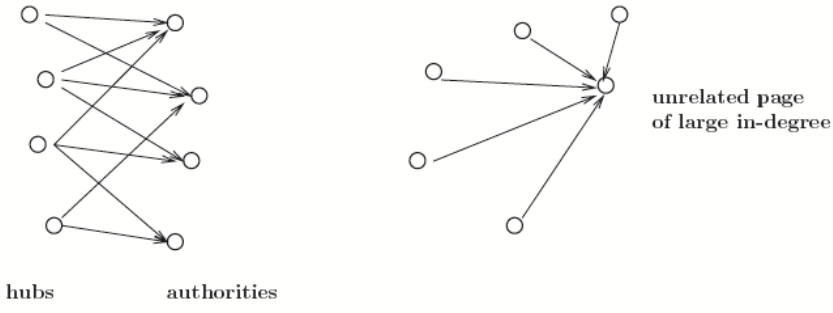

Figura: 2.8: A distribuição entre Hubs e Authority segundo Kleinberg

Entretanto, em vez de qualificar uma página simplesmente como 'hub' ou 'authority' (que é mais ou menos o que foi proposto no algoritmo Hilltop), o algoritmo de Kleinberg atribui a cada página dois índices, um 'hub index' e um 'authority index'. O peso de cada link dependerá dos índices hub e authority da página em que se encontra. O processo de cálculo é recursivo e pode envolver biliões de páginas. Durante a sua concepção, o algoritmo mostrou-se impraticável porque exigia volumes enormes de recursos computacionais.

2.2.7 As várias gerações de motores de pesquisa

Temos vindo a verificar da necessidade de dar resposta e de resolver alguns problemas relacionados com a primeira preocupação do modelo Internet: **de absorver** [Marcondes & Sayão, 2002] **e suportar todo o tipo de informação e disponibilizá-la para toda a gente**. O 'emaranhado' de nós interconectados [Schons, 2007], desprovidos de qualquer organização, proporcionaram o desenvolvimento de tecnologia visando a pesquisa e recuperação de informação de forma eficiente do ponto de vista do utilizador.

Como vimos, os algoritmos de Ranking (PageRank e HITS) marcaram o desenvolvimento de uma nova geração de motores de pesquisa, que pretendiam dar uma resposta ordenada às solicitações dos utilizadores. Infelizmente, uma vez mais os spammers têm encontrado formas de contorná-los e incluir respostas indesejáveis em relação à pesquisa pretendida.

No modelo do PageRank uma página tem o seu próprio valor: a sua classificação não é condicionada por nenhum assunto em particular. Então os Spammers implantaram sites com conhecimentos sobre assuntos irrelevantes e conseguem adquirir alta classificação nos seus sites especializados. Eles utilizam a técnica já referida de 'bandwagon' para a sua rede de sites (vizinhança), criando o que se pode chamar de 'sociedade de admiração mútua' (MAS). Trata-se da conhecida técnica publicitária de depoimentos [Lee & Lee(eds.), 1939; Metaxas & DeStefano, 2005], por exemplo de figuras públicas (animadores, futebolitas, artistas de TV, etc), que transmitem a sua opinião sobre questões fora da sua área de especialização.

Por esta última razão, o modelo HITS mostrou-se muito permissivo

aos spammers, sobretudo devido ao facto de a sua eficácia depender da precisão do cálculo inicial de vizinhança.

A tabela 2.1 (Fonte: [Metaxas & DeStefano, 2005]) resume algumas pesquisas sobre as primeiras gerações de motores de pesquisa e os principais modelos de ataques que sofreram, agrupadas pelo autor em 2005.

SE	**Ranking**	**Spamming**	**Propaganda**
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation	+ mutual admiration	+ testimonials societies

Tabela: 2.1: Web Spam, Propaganda and Trust

Esta 'cultura humana de produção' referida por [Schons, 2007], não-hierárquica e interactiva (pelo seu alto grau de interactividade), promove a remodelação na estrutura do fluxo de informação. Nessa perspectiva, o mesmo autor, reflecte acerca da contribuição entre os internautas e atribui-lhes o termo 'inteligência colectiva', porque todos podem contribuir para a concretização de uma 'tecnodemocracia' por intermédio das suas percepções e inteligências.

Hoeschl [Hoeschl, 2006] apresenta outra forma de classificação. Trata-se de uma **classificação temporal baseada em gerações** quanto aos mecanismos de busca na Web, diferente da apresentada por Metaxas.

- A primeira geração, para fazer frente ao grande volume de conteúdo na Internet, deu-se com os directórios ou catálogos (Yahoo e similares), e a descoberta, avaliação, descrição e inclusão dos recursos eram feitas por profissionais de informação.
- Na segunda geração esse processo foi automatizado com os robots digitais (Altavista).
- A terceira geração surgiu com os metabuscadores, juntando num único resultado as informações de vários motores de busca (MetaMiner).
- Na quarta geração os resultados são ainda mais refinados (All the Web).

- A quinta geração corresponde à geração actual, tendo como grande exemplo o Google, utilizando várias tecnologias de evolução dos algoritmos de Ranking.

  Baseado na classificação das ferramentas de busca proposta por Hoeschl [8], a sexta e sétima geração ainda se encontram em período de desenvolvimento e testes.

- A sexta geração compreende a junção de vários tipos de arquivos diferentes em um mesmo processo de busca (tecnologia A9) [9].

- Já a sétima geração de ferramentas de pesquisa traz consigo uma tecnologia inovadora, baseada em ontobuscadores, conjugando tecnologias inteligentes com conhecimentos milenares e filosóficos, penetrando na essência dos conceitos e objectos.

A evolução desta ultima fase, que melhor analisaremos na secção seguinte, inclui a análise semântica que, segundo Souza [Souza & Alvarenga, 2004], representa a evolução da web actual baseada em documentos de hiper texto escritos segundo a linguagem HTML, que só permitem a indexação automática por palavras chaves, extraídas do texto.

2.3 O que se espera dos motores de busca

Em face das evoluções referidas, mas também das novas necessidades dos utilizadores, os especialistas em motores e técnicas de RI estão agora a desenvolver esforços no sentido de encontrarem uma nova geração de motores de pesquisa [Broder, 2002; Metaxas & DeStefano, 2005], que consigam interpretar as necessidades do utilizador 'por detrás da consulta', usando técnicas de interpretação semântica das queries digitadas pelo utilizador, já de uma forma mínima, normalmente como sugestão, disponibilizada por alguns motores.

Para Marcondes [Marcondes, 2007], a infra-estrutura da web semântica consiste em páginas utilizando XML, que, além do conteúdo, terão metadados utilizando vocabulários e relações muito poderosas (ontologias e RDF - Resource Description Framework) para expressar a semântica das novas páginas Web. Nesse ambiente, actuarão agentes inteli-

[8] 2006
[9] http://www.a9.com

gentes que irão realizar tarefas que envolvem conhecimento, raciocínio e dedução.

A abordagem do uso de tecnologias inteligentes enquanto elemento potenciador para incrementar o processo de revolução digital e informacional tem sido alvo de importantes estudos. A ideia de tornar a web capaz de aprender (armazenar, recuperar e processar informações) de forma inteligente, similarmente a um grande cérebro global, tem vindo a ganhar consistência. Oliveira e Vidotti [Oliveira & Vidotti, 2004] defendem que, para a formação de uma inteligência colectiva mais dinâmica é fundamental que a própria rede descubra e aprenda a melhor organização para si mesma.

Considerando-se como parâmetro a mente humana, o conhecimento e significado decorrem de um processo de aprendizagem em que, quanto maior o uso de determinados conceitos, mais fortemente eles se ligam. Para a web a análise é semelhante: **com base nos caminhos mais percorridos pelos internautas, algumas conexões tornam-se mais importantes, enquanto os links pouco utilizados tornam-se menos importantes**.

Um exemplo da tecnologia dos ontobuscadores é o Onto-Web, um buscador inteligente que é baseado em ontologias e técnicas de inteligência artificial, capaz de 'pensar' enquanto seleciona as informações. O seu grande diferencial é que, conforme Hoeschl [Hoeschl, 2006], utiliza semânticas e estruturas valorativas para contextualizar as buscas e refinar resultados. Além disso, o seu motor de busca, efectua a hierarquização de conteúdos baseando-se em métricas de similaridade e engenharia do conhecimento. O mesmo autor refere que, além das ontologias, o sistema utiliza diversas outras tecnologias como PCE (Pesquisa Contextual Estruturada), RC2D (Representação do Conhecimento Contextualizado Dinamicamente), técnicas de mineração de dados e raciocínio baseado em casos.

Outro grande diferencial deste motor é a sua facilidade para comparar textos, pois enquanto outros pesquisadores como o Google aceitam até 256 caracteres, o OntoWeb permite entradas de até 7000 caracteres. Até ao momento, a ferramenta possui uso limitado[10], mas futuramente pretende-se utilizá-la como um motor de pesquisa comum para todos os tipos de assunto.

[10] Ex^{o}.:governo eletrónico

A adopção de um conjunto de tecnologias inteligentes que fazem uso de semânticas, ontologias, redes neurais e inteligência artificial, parece constituir a base de elementos essenciais no futuro para que o processo de aprendizagem se torne, de facto, realidade na grande rede, tornando viável a sua reestruturação.

Nessa perspectiva futura de construção da Internet como um espaço voltado para a aprendizagem, apontamentos têm sido realizados acerca do surgimento de uma nova era das redes, intitulada Web 3.0. Esta nova rede poderá vira a actuar como um especialista, respondendo a perguntas dos utilizadores a partir da sua própria análise.

Desse modo, os utilizadores não precisariam efectuar longas pesquisas para emitir conclusões porque a própria rede encontraria as melhores soluções.

A Web 3.0, enquanto protótipo, tem como premissa levar em conta o sentido de cada palavra do utilizador, efectuando ligações entre elas para que o resultado seja preciso conforme o contexto do utilizador.

Ela actuará com inteligência e intuição. Sobre esta nova fase da web, que será dotada da capacidade de 'aprender', 'raciocinar' e 'entender', [Johnson, 2003] aponta que:

> Pela impossibilidade de a web vir a ser semelhante à consciência humana, não podemos aferir que seja incapaz de aprender. Antes pelo contrário: 'Uma rede de informação adaptável, capaz de reconhecer padrões complexos, poderá vir a ser uma das invenções mais importantes de toda a história da humanidade'.

Para que os motores de pesquisa possam responder às necessidades crescentes da procura, é necessário que respondam satisfatoriamente, quer em velocidade de resposta, quer em qualidade da informação. Neste sentido há necessidade de se verificar um compromisso elevado entre as componentes de hardware e de software, que poderemos resumir em alguns dos pontos seguintes.

2.3.1 Scalability

Escalabilidade é uma propriedade fundamental para qualquer sistema. Sendo difícil de definir por palavras, é o indicador da sua capacidade (entenda-se: do sistema) para lidar com quantidades crescentes de trabalho de maneira simples e transparente para o utilizador.

Um grande indexador deve ser escalável [Risvik & Michelsen, 2002] tanto no que diz respeito à capacidade de armazenamento de documentos como quanto à capacidade de recuperação de documentos.

São vários os indicadores que contribuem para a classificação positiva do indicador de escalabilidade:

- **Crawling**

  Por definição, como referido por Risvik [Risvik & Michelsen, 2002], 'Crawler' refere-se a um programa (ou módulo) que agrega dados no ambiente WWW de forma a torná-los pesquisáveis. Embora utilizando diversas técnicas heurísticas e diversos algoritmos, podemos dizer que a maioria deles são baseados na observância das sequência de links.

  Segundo o mesmo autor cada máquina de 'crawling' é responsável por toda a recuperação, processamento e armazenamento dos documentos. As diferentes máquinas, constituindo um cluster de rastreio, podem trabalhar independentemente, excepto para a troca de URLs novos.

  Devido a essa independência, a capacidade de armazenamento, C_S, e a capacidade de tratamento, P_b, são a soma da capacidades individuais de cada máquina:

$$C_S = \sum_i C_{S,i} \tag{2.10}$$

$$C_P = \sum_i C_{P,i} \tag{2.11}$$

  Também conhecidos por wanderers, web robots, aranhas, web spiders ou worms[Kobayashi & Takeda, 2000], os 'crawlers' evocam o imaginário, o intangível, da Web e são necessariamente muito rápidos [Levene & Poulovassilis, 2004; Pant *et al.*, 2004].

Segundo a Wikipedia, este processo de rastreamento da web ou *spidering*, proporciona a actualização de dados referentes, usualmente, a cópias de todas as páginas visitadas para posterior tratamento por um motor de busca. Destas serão criados indíces para fornecer pesquisas mais rápidas.

Em geral, o 'crawler' ou 'robot', começa com uma lista de URLs para visitar, chamada de sementes. Ao percorrer estas URLs identifica todos os hiperlinks na página e adiciona-os à lista de URLs para visitar, no que é conhecido como rastreamento fronteira. Os URLs da fronteira são recursivamente visitados, de acordo com um conjunto de políticas, baseadas essencialmente no grande volume de dados, no rápido ritmo de mudança e na criação dinâmica de páginas.

O grande volume de dados implica que o indexador só possa baixar algumas das páginas do universo total em tempo útil, por isso ele precisa de organizar os downloads. Por outro lado, a elevada taxa de alteração dos conteúdos das páginas implica que, no momento em que o indexador está a executar o download da última página de um site, seja muito provável que novas páginas tenham sido acrescentadas ao site, ou ainda páginas que já tenham sido actualizadas ou mesmo suprimidas.

Acresce ainda que - a criação dinâmica de páginas - proporciona uma infinidade de combinações sem conteúdo novo. Devido à multiplicidade de referências usadas para o mesmo objecto, apenas alguns dos links trarão informação exclusiva. Um exemplo simples pode ser demonstrado pelas inúmeras possibilidades de endereçar uma fotografia: pelas miniaturas, pelos diversos tipos de imagens associados e, acima de tudo, pela quantidade de diferentes URLs que podem apontar para a mesma fotografia. Calcular as combinações possíveis e detectá-las de forma a guardar apenas uma instância - e isso é tratar o que é verdadeiramente exclusivo - é um sério problema, também matemático, para resolver pelos indexadores. Por isso Edwards et al. [Edwards *et al.*, 2001] diz que o 'crawler' deve escolher cuidadosamente qual a próxima URL a visitar.

Quanto à construção do software um 'crawler' deve juntar a essa política de selecção uma arquitectura que permita elevadas performances. Porque este software é o 'core' dos motores de pesquisa,

os seus algoritmos e arquitecturas são muito protegidos do conhecimento público. Esse secretismo tem essencialmente a ver com a construção de contra-medidas principalmente pelos especialistas na criação de *spamdexing*, que encontram sempre incentivos para a batota.

Cho et al. [Cho *et al.*, 1998] estudou qual a relevância de um crawler visitar frequentemente as páginas rastreadas, de forma a procurar páginas 'importantes'. O estudo conclui que a utilização de um bom 'ordering schema' como auxílio à interpretação do crawler ajuda muito na decisão de actualização, aumentando significativamente a performance de que falávamos.

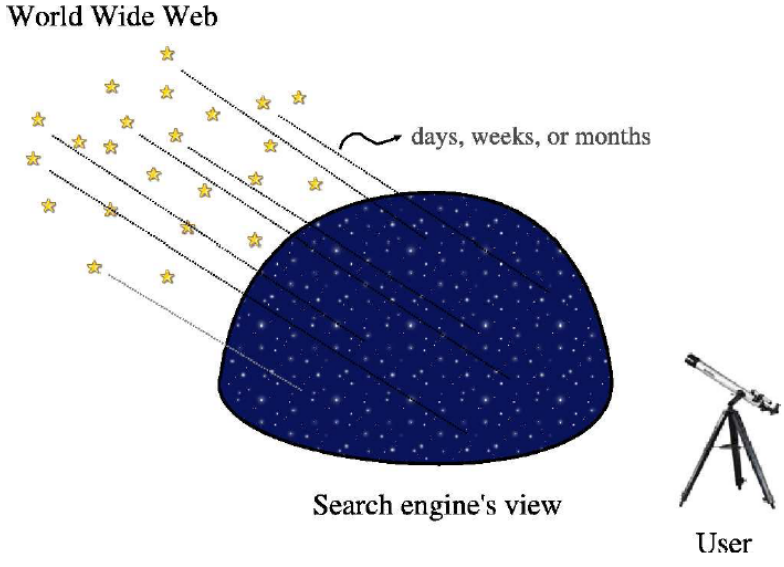

Figura: 2.9: Rastreando a Web infinita

Uma ideia e imagem esclarecedora é referida por Castilho [Castillo, 2004] ao referir que:

> Crawling a Web, de certo modo, assemelha-se ao olharmos para o céu numa noite clara: o que reflecte o estado de ver as estrelas em momentos diferentes, enquanto a luz percorre diferentes distâncias. As últimas páginas que estão a ser classificadas provavelmente são representadas com muita precisão, mas as primeiras páginas que foram

guardadas têm uma elevada probabilidade de terem sido alterados. Esta ideia está representado na figura 2.9.

- **Indexing**

  Como definição refere Risvik [Risvik & Michelsen, 2002] que 'indexer' é um módulo, dentro do conceito referencial de motor de busca (Fig. 2.10) que tem uma colecção de documentos ou dados e constrói um índice pesquisável a partir deles. Práticas comuns são arquivos invertidos, espaços vectoriais, estruturas e híbridos destes.

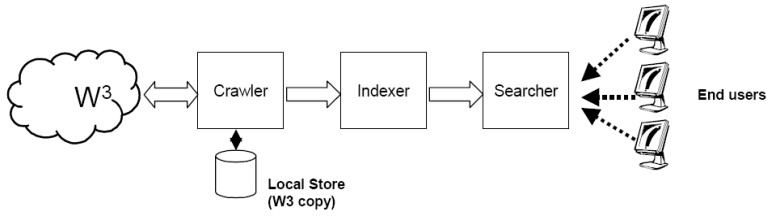

Figura: 2.10: Modelo de Motor de busca

- **Searching**

  Segundo o mesmo autor [Risvik & Michelsen, 2002] O 'searcher' trabalha ao nível do indexador, pormenoriza que trabalha *à saída* do indexador. Aceita as consultas do utilizador, executa-as ao longo do índice, e retorna os resultados à entidade que pergunta.

- **Ranking**

  Infelizmente, os motores de busca não têm a capacidade de fazer algumas perguntas, como faz um bibliotecário, para centrar a pesquisa. Também não podem invocar experiências anteriores para classificar as páginas da web, na forma como nós, seres humanos, podemos.

  Então os indexadores seguem um conjunto de regras, conhecidas como algoritmos de ranking, cujo objectivo é responder à pergunta que originou a pesquisa com uma lista de apontadores para sítios que possam dar resposta a essa pergunta. Esta lista é ordenada segundo os algoritmos de ranking, que têm evoluído com o tempo e

com a necessidade de ultrapassar constrangimentos que vão sendo detectados.

No início deste capítulo referimos já, com algum detalhe, alguns algoritmos de ranking mais usados.

2.3.2 Relevance

Depois de muitos estudos realizados na área de Recuperação de Informações, de acordo com Buckley [Buckley & Salton, 1995], os pesquisadores concluíram que o utilizador não consegue recuperar os documentos que realmente precisa na primeira vez que efectua uma consulta ao sistema. Tipicamente o que acontece é que o utilizador efectua uma consulta inicial como primeira tentativa, de certo modo exploratória que vai refinando, melhorando-a, conforme os resultados que vai obtendo.

As próximas consultas passam a recuperar cada vez mais documentos mais relevantes para o utilizador, pois ele vai contextualizando mais precisamente o assunto que é de seu interesse [Agichtein *et al.*, 2006], utilizando novas palavras e eliminando as palavras que prejudicam as suas consulta, obtendo documentos que estão fora de seu interesse. Desta forma são produzidas novas consultas que teoricamente serão mais precisas e reflectindo as suas necessidades, recuperando mais documentos relevantes.

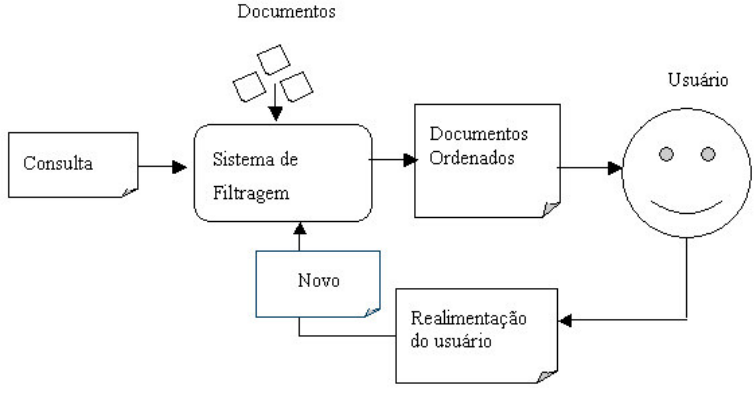

Figura: 2.11: Método do Relevance Feedback

Esta é uma ideia polémica devido ao carácter subjectivo do julgamento da relevância de documentos.

A tarefa de analisar a relevância de vários documentos perante um tópico é um tanto ou quanto árdua, porém possível quando a colecção possui poucos documentos. No entanto, a pesquisa em todo o ambiente web tornou impossível a tarefa de analisar os documentos um a um.

Assim, segundo o mesmo autor [dos Santos Batista Junior, 2006], a fim de poder utilizar um grande número de documentos em avaliações sem a necessidade dessa análise individual, surgiu a idéia de pooling, que consiste numa **estratégia para efectuar julgamentos de relevância de maneira não exaustiva**, já que somente um conjunto de documentos recuperados por um conjunto de sistemas é analisado. Os documentos que não estiverem nesse conjunto são considerados irrelevantes sem passar por qualquer julgamento.

Uma melhoria no conceito de relevância, que inclui na análise o feedback do utilizador, é defendido por Hiemstra [Hiemstra & Robertson, 2001]. Segundo esse estudo, relevância implica reponderação dos termos de consulta baseados em alguns feedback do utilizador, de forma explícita ou implícita. Considera como feedback, para o conceito de relevância, todas as acções produzidas pelo utilizador em documentos anteriormente recuperados para a construção de uma consulta.

Nesse estudo [Hiemstra & Robertson, 2001] são referidos dois modelos probabilísticos de recuperação de informação: o modelo de linguagem natural (language models) e o segundo baseado no modelo probabilístico da independência binária (binary independence model). O documento mostra a semelhança das abordagens na relevância do feedback destes modelos, e avalia a pertinência de novos algoritmos de feedback.

- O modelo 'language model for relevance feedback':

$$P(D).P(T_1, T_2, ..., T_n|D) = P(D).\prod_{i=1}^{n}((1-\lambda_i)P(T_i) + \lambda_i P(T_i|D)) \tag{2.12}$$

Havendo vários modelos similares a este apresentados por [Miller *et al.*, 1999], [Kg, 1999] e [Ponte & Croft, 1998], com a excepção

de [Ponte & Croft, 1998] todos usam a interpolação linear de probabilidades.

- O modelo binário de independência probabilística 'The binary independence probabilistic model':

  Este modelo binário afirma que, dada uma relevância L (e irrelevância $\bar{L}$), os atributos Ai de um documento D são estatisticamente independentes.

  No caso de recuperação de texto, os atributos do documento são simplesmente os termos do próprio documento. O valor do atributo é 1, quando o termo a pesquisar está presente no documento, ou 0 se está ausente.

  Conforme [Hiemstra & Robertson, 2001; Robertson & Jones, 1976], simbolicamente, o modelo binário de independência probabilística, apresenta-se com a seguinte descrição:

$$O(L|D) = O(L).\prod_{i=1}^{l}\frac{P(A_i|L)}{P(A_i|\bar{L})} \tag{2.13}$$

  Na fórmula l é o número de termos diferentes na consulta, e $O(L|D)$ é a probabilidade de relevância de um documento:

$$O(L|D) = P(L|D)/(1 - P(D)) \tag{2.14}$$

As experiências de feedback relevante relatadas por Hiemstra [Hiemstra & Robertson, 2001] estão de acordo com as experiências descritas por Robertson e Sparck-Jones [Robertson & Jones, 1976] e, mais recentemente, por Sparck-Jones et al. [Sparck-Jones *et al.*, 2000].

Nessas experiências, ao sistema é apresentada a lista completa de todos os documentos pertinentes. Isto é, de facto, um cenário irrealista: Na prática, uma tarefa preditiva de relevância da ponderação, onde ao sistema são apresentados apenas alguns dos documentos relevante para a lista, será muito mais interessante.

Se todas as informações relevantes forem postas à disposição do sistema, então, esperamos que o sistema atinja um desempenho ideal. Naturalmente, esperamos que o sistema nunca sirva para diminuir o desempenho de uma consulta.

2.3.3 Static Ranking

- **Content Quality**

  Até à data a maior parte do trabalho no cálculo do PR de uma página da Web centrou-se sobre a melhoria da ordenação dos resultados retornados para o utilizador, logo dependentes do query inicial (*dynamic ranking*). No entanto, Richardson [Richardson *et al.*, 2006] refere que uma boa classificação independente (*static ranking*) também é crucialmente importante para um motor de busca.

  Desde a publicação de Brin e Page do documento sobre PageRank, a ordenação das páginas na comunidade da Web têm dependido do algoritmo de PageRank[11].

  No entanto outros trabalhos de investigação têm avançado no sentido de melhorar esses algoritmos de classificação, como é o caso do Hilltop em relação ao PR original.

  No trabalho de Richardson [Richardson *et al.*, 2006] mostra que podemos superar significativamente o PageRank usando recursos que são independentes da estrutura de links da web, usando um novo algoritmo de RankNet. Pretende nesse trabalho demonstrar que se ganha **um novo impulso na precisão**, utilizando os dados sobre a frequência com que os utilizadores visitam as páginas na web.

  O mesmo algoritmo de RankNet é classificado por [Svore *et al.*, 2007a] como sendo um algoritmo de base rede neural utilizado para classificar um conjunto de factores de produção, neste caso, o conjunto de frases num determinado documento.

[11] Já referidos neste capítulo, na secção 2.2.3, bem como outros tipos de algoritmos de ranking.

Capítulo 3

Visibilidade dos sites

Search Engine Optimization (SEO) é simultaneamente uma arte e uma ciência. [1]

Utilizando a definição de Perkins [Perkins, 2001] um motor de pesquisa usa automatismos, tais como robots (também conhecidos por spiders, aranhas, etc.) e indexadores, para criar índices na Web. Permite a pesquisa desses índices mediante determinados critérios de busca e devolve um conjunto de resultados ordenados pelo grau de relevância dos critérios da pesquisa.

Podem considerar-se duas áreas distintas, num estudo mais analítico como o de Baeza-Yates [Baeza-Yates *et al.*, 2007b], que identifica duas partes nos motores de pesquisa: A primeira parte onde evoluem o crawler e o indexer; e uma segunda parte, tratada on-line, destinada a processar os queries de pesquisa e a produzir a lista de resultados.

Como iremos ver a visibilidade dos sites, representada na lista de resultados ordenada 'pelo grau de relevância', é muito disputada.

3.1 Introdução

Como já citámos, os autores [Gyongy *et al.*, 2004; Gyongyi & Garcia-Molina, 2005; Svore *et al.*, 2007b] definem como principal característica

[1] http://www.free-ebooks.net/ebook/apr07/MYSEOEGUIDE.pdf

do Web-Spam a capacidade de iludir os motores de pesquisa, no momento da atribuição do ranking de representatividade, atribuindo às páginas valor superior ao merecido, aumentando assim o seu grau de visibilidade para potenciais visitantes.

Esta ilicitude gerou um 'braço de ferro' [Castilho *et al.*, 2006; Sydow *et al.*, 2008] entre os administradores dos motores de pesquisa, que tentam manter estável a credibilidade que criaram junto dos utilizadores, contra todos os que usam técnicas designadas por 'black hat', que manipulam a ordem natural do ranking.

Devido ao alto valor e segmentação dos resultados da pesquisa, há um relacionamento contraditório entre os motores de pesquisa e SEO's[2] (Optimizadores de motores de pesquisa).

Na tentativa de minimizar este fosso de incompreensão e os efeitos agressivos e prejudiciais aos fornecedores de conteúdos, em 2005 foi realizada a primeira conferência anual designada por AirWeb.

Com efeito, tal como mencionado por [Becchetti *et al.*, to appear (In Press)], existe uma grande área cinzenta entre atitude 'ética', defendida pelos detentores de SEO's e a opinião de administradores dos motores de pesquisa que classificam esta atitude de spam 'antiético' [Svore *et al.*, 2007b].

Mbikiwa and Weideman [Mbikiwa & Weideman, 2006] referem que:

> White hat techniques are considered to be ethical and above board, as viewed by a search engine crawler [Weideman, 2007]. These include judicial keyword placement, correct use of metatags, single submissions to search engines and avoiding the use of frames. Black hat techniques are considered to be unethical, since they attempt to present a website in such a way to a crawler, that the website earns a higher ranking than what it deserves by virtue of its contents.

De facto a função dos SEO's é a de fornecer uma gama de serviços extraordinários que pretendem assegurar que as páginas em que intervêm são indexáveis pelos indexadores, para a criação de milhares ou milhões de páginas falsas, garantindo uma boa posição nessa indexação. Aos

[2] Acrónimo do Inglês 'Search Engine Optimization'

algoritmos de Ranking cabe agora a árdua tarefa adicional de detectar essas páginas usando técnicas mais elaboradas [Krause *et al.*, 2008].

3.2 Técnicas de 'Search Engine Optimization' - SEO

A expressão 'optimização de sites para a pesquisa', refere-se a um conjunto de estratégias com o objectivo de melhorar a posição nos resultados naturais (orgânicos) nos sites de busca[3].

Segundo Weideman [Weideman, 2007], SEO refere-se a todo o procedimento que envolva a alteração (ou a preparação inicial) de uma página com o destino de ser 'crawler-friendly', no sentido de permitir uma indexação mais rápida pelos crawlers.

As técnicas utilizadas pelos SEO, intervindo ao nível interno do Web site, têm como finalidade adaptar funcionalidades que melhor respondam a uma grande variedade de palavras chave relevantes ao seu conteúdo. Com isto promove-se potencialmente a resposta aos motores de pesquisa e daí a sua melhor posição final. Estratégias de SEO podem melhorar tanto o número de visitas quanto a qualidade dos visitantes, onde qualidade significa que os visitantes terminam a acção esperada pelo proprietário do site (ex. comprar, assinar, aprender algo).

Das variadas técnicas existentes com este objectivo, muitas baseiam-se nos algoritmos dos motores de pesquisa - 'colocam as peças no lugar certo'.

Alguns objectivos relacionam-se com o querer atingir todo o tipo de tráfego na rede, donde os sites podem ser optimizados para incrementar a busca de frases comuns. Uma boa estratégia para optimização nos sistemas de pesquisa pode funcionar perfeitamente com sites que tenham interesse em atingir um grande público-alvo, tais como Informativos Periódicos, serviços de directórios, guias, ou sites que exibem publicidade com um modelo baseado em CPM.

Em contraste muitas empresas tentam optimizar os seus sites para um grande número de palavras-chave altamente específicas que indicam uma maior preparação para venda. Optimizações deste tipo, tendo em

[3]http://pt.wikipedia.org/wiki/Otimização para Sistemas de Busca

vista um amplo espectro de termos para busca, pode impedir a venda de produtos por gerar um grande volume de solicitações com baixa qualidade, com um custo financeiro alto e resultando em pequeno volume de vendas.

Chambers [Chambers, 2006] descreveu um modelo, baseado em experiências concretas, para melhorar a visibilidade de um site, dentro uma determinada categoria de sites.

Este modelo enumera uma quantidade de factores preponderantes no que temos vindo a designar por visibilidade de um site, conforme as figuras 3.1 e uma evolução estudada por Visser [Visser *et al.*, 2007] 3.2.

NUMBER	LEADING VISIBILITY ELEMENTS	RANK
1	Inclusion of meta-tags	1.5
2	Hypertext / Anchor text	2
3	No Flash or fewer than 50% of content	3
4	No Visible Link Spamming	4
5	Prominent Link Popularity	4.5
6	No Frames	5
7	Prominent Domain Names	7
8	Prominent Headings	7
9	No Banner Advertising	8
10	Prominent HTML Naming conventions	10

Figura: 3.1: Modelo de Chambers sobre a visibilidade de um website.

Nesse seu estudo, Visser [Visser *et al.*, 2007] expandiu o modelo original de Chambers para efectuar a distinção entre elementos visíveis que devem ser incluídos ('Essentials'), que podem ser incluídos ('Extras'), os que devem ser evitados ('Cautions') e os que não devem ser utilizados ('Dangers') [Weideman, 2007].

O termo SEO também se refere à indústria de consultoria, que trabalha na optimização de projectos e de websites. Neste grupo de profissionais podemos considerar dois grupos de intenções: '**SEO de White Hat**' (geralmente utilizam métodos aprovados pelos sistemas de pesquisa, como a prática de construção de conteúdo e melhoria da qualidade do site), e os '**SEO de Black Hat**' (utilizam truques como 'cloaking' - camuflagem do conteúdo real da página - e spamdexing).

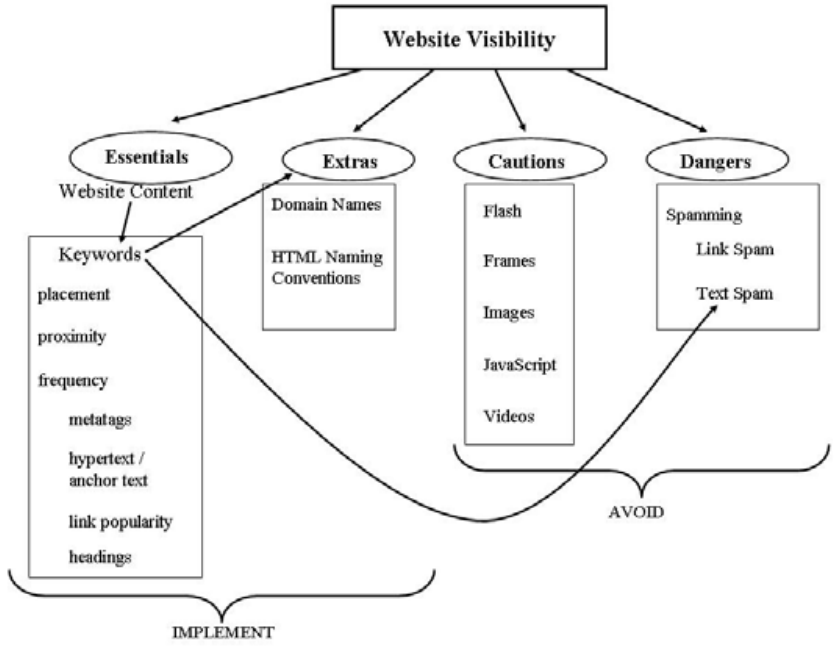

Figura: 3.2: O Modelo sobre a Visibilidade de um website, segundo Visser

Os praticantes de técnicas de White Hat dizem que os Black Hat se valem de métodos que são vistos como tentativas de manipular o posicionamento nos resultados dos motores de pesquisa a seu próprio favor. Os adeptos do Black Hat dizem que todas as tentativas e técnicas utilizadas para manipular o posicionamento são legítimas, e o tipo de técnica utilizada para se alcançar o objectivo de melhorar o rank é irrelevante.

Podendo ser incorporado como um serviço, as técnicas de SEO produzem mais resultado se incorporadas no projecto inicial de desenvolvimento do site.

As motivações para a melhoria de resposta ao sistema de pesquisa divergem muito consoante a especificidade dos objectivos individuais.

3.2.1 Histórico

3.2.1.1 Início dos sistemas de pesquisa

Alguns webmasters e responsáveis dos portais de conteúdo começaram a optimizar os seus sites para os sistemas de pesquisa em meados dos anos 90, porque os primeiros motores de busca estavam a introduzir o conceito de indexação do conteúdo na WWW. No início, os webmasters precisavam enviar o endereço do site aos vários sistemas de pesquisa existentes na Rede, para que programas do tipo aranhas[4], pudessem 'mapear' o site e armazenar as informações recolhidas. O padrão e suporte dos pesquisadores era mapear uma página da web inteira e seleccionar as palavras relacionadas na busca; então uma página com muitas palavras diferentes ampliava a combinação de resultados e uma página da web, contendo uma relação de palavras como um dicionário, teria como resultado um grande número de combinações, consequentemente limitando os resultados somente a nomes únicos.

Os sistemas de busca classificavam a informação por tópicos, muitas vezes exibindo como resultado das buscas páginas já expiradas, não existentes ou de conteúdo diferente ao que estava armazenado. Com o número crescente de documentos online e vários webmasters trabalhando para aumentar o valor nos resultados em busca orgânica, os sistemas de buscas mais populares começaram a classificar as páginas de resultados mais relevantes em primeiro. Este foi o início de uma fricção desconfortável entre Sistemas de Busca e Webmasters que continua até hoje.

Os primeiros motores de busca eram orientados pelos próprios webmasters. Na ocasião, as versões existentes dos algoritmos utilizados nos sistemas de busca confiaram aos webmasters e provedores de conteúdo a responsabilidade no fornecimento das informações na forma de Categorias e o uso das Meta-Tag para palavras-chave (meta tags ou sistemas de busca que usavam arquivos de índice como o ALIWEB), fornecendo assim um guia para o índice de cada página. Quando alguns webmasters começaram a abusar no uso das Meta-Tags, fazendo com que as informações aí contidas não correspondessem ao conteúdo, os motores de busca abandonaram esta forma de obter informações através de Meta-Tags e desenvolveram um sistema de ranking mais complexo utilizando algoritmos, elevando a filtragem das palavras e elevando o número limitado

[4]Spiders ou crawlers

para palavras (anti-dicionário), incluindo:

`Texto dentro do Tag de título`

`Nome de Domínio`

`URL` (de Universal Resource Locator) em português significa (Localizador Uniforme de Recursos) directórios e nomes de arquivos HTML element

`Keyword density - Densidade das palavras`

`Proximidade das Palavras-Chave`

`Atributos do texto alternativo para imagens`

`Textos dentro da Tag NOFRAMES`

Pringle et al. [Pringle *et al.*, 1998], também definiu um número de atributos, dentro da codificação HTML de uma página, que frequentemente eram manipulados por provedores de conteúdo na Internet tentando melhorar a própria classificação em motores de busca. Devido a factores que estão praticamente sob o controle exclusivo dos responsáveis por um site, os motores de busca continuaram a sofrer abusos e tentativas de classificações manipuladas. Daí, e no sentido de fornecer melhores resultados aos seus utilizadores, os sistemas de busca tiveram que se adaptar e assegurar que as páginas de resultados mostrassem sempre os resultados mais relevantes durante uma pesquisa, ao invés de páginas inúteis cheias de palavras-chave e termos criados de forma pouco escrupulosa por webmasters, na tentativa de as usar como iscas para exibirem webpages sem link ou conteúdo. Estes factos levaram ao nascimento de um novo tipo de motor de busca.

3.2.1.2 Sistemas de Pesquisa Orgânica

O pesquisador Google foi iniciado por dois estudantes de PhD da Universidade de Stanford: Sergey Brin e Larry Page e trouxeram um novo conceito para avaliar páginas da web. Este conceito, chamado **PageRank**, e já referido na secção [2.2.3], foi determinante para o início da criação dos algoritmos do Google.

O PageRank trabalha principalmente com o factor link e usa a lógica deste sistema de ligação entre páginas e, de uma forma simplista, atribui

o valor de um voto para a página em questão. O facto de uma página ser referenciada por um link, funciona de forma a validar a existência do site e dar um valor mais 'digno' ao voto.

Com ajuda do PageRank, o Google provou ser muito eficiente em oferecer respostas relevantes nas páginas de resultados e tornou-se o motor de pesquisa melhor sucedido e mais popular.

Em virtude de o sistema de PageRank fazer a avaliação de websites através de factores externos, ou fora do controle de pessoas em particular, o Google sentiu que desta maneira poderia ser mais difícil manipular a relevância de uma página.

No entanto alguns webmasters já tinham desenvolvido ferramentas para manipulação de link´s e esquemas para influenciar o sistema de pesquisa Inktomi. Estes métodos provaram ser igualmente aplicáveis aos Algoritmos do Google. Muitos sites focaram as suas acções em trocar, comprar e vender links numa escala imensa. A confiança do PageRank no sistema de link como um **voto de confiança para um valor da página foi subvertido** dado que muitos webmasters procuraram vender links simplesmente para influenciar o Google em enviar mais tráfego, independentemente dos links serem realmente útil aos visitantes humanos do site.

Para complicar a situação, *o suporte de pesquisa* omitia o mapeamento de uma página web inteira para procurar as chamadas palavras relacionadas nas páginas web, e uma página web contendo uma listagem do tipo dicionário ainda combinaria quase todas as pesquisas (excepto nomes especiais), alcançando um link-rank mais alto. Sobressai aqui o facto de que páginas de dicionários e links para esquemas muito trabalhados poderem fazer distorcer dramaticamente os resultados.

Importa também considerar que a Internet continua a chegar a uma grande quantidade de utilizadores não técnicos, que geralmente não conhecem técnicas avançadas de pesquisa que os ajudem a encontrar a informação que podem estar a pesquisar. Além disso o volume e complexidade dos dados indexados tiveram um enorme crescimento em relação aos primórdios da Internet.

Para classificar os sites, os mecanismos de pesquisa levam em consideração os seguintes aspectos:

- Idade do sítio (site)
- Há quanto tempo o domínio está registado
- Idade do conteúdo
- Frequência do conteúdo: regularidade da actualização do conteúdo
- Tamanho do texto: número de palavras acima de 200-250 (não afectava o Google em 2005)
- Idade do link e reputação do sítio que o aponta (outbinds)
- Standardização dos factores on-site (meta-tags comuns sem evidência de manipulação)
- Pontuação negativa para factores on-site (por exemplo, uma redução da pontuação de websites com extensas palavra-chave de meta-tags por ser um indicativo de ter sido otimizada por SEO)
- Conteúdos únicos
- Relacionamento dos termos utilizados com o conteúdo (os termos associados ao motor de pesquisa como estando relacionadas com o conteúdo principal da página)
- Google Pagerank (Exclusivo para os motores que usam o Algoritmo do Google)
- Ligações externos, o texto âncora contido nesses links externos e nos sites/páginas que os contenham
- Citações e fontes de investigação (a indicação do conteúdo pode classificar a investigação como de qualidade)
- Termos com relacionamento morfológico (finanças / financeiro / financiamento)
- Links a apontarem para o site (incoming links) e o texto âncora dessas referências
- Pontuação negativa para alguns incoming backlinks (p.ex. venham de páginas de baixo valor por mera retribuição de links, etc)

- Taxa de aquisição de backlinks: demasiadas ou demasiado rápido, poderá indicar links 'não naturais' ou seja pode representar que houve compra de links apenas para aumentos de ranking
- Texto envolvendo ligações externas recebidas e backlinks. Um link na sequência das palavras 'Sponsored Links' pode ser ignorado
- Utilização de 'rel=nofollow' a sugerir que o motor de busca deve ignorar a ligação
- Profundidade do documento no site
- Métricas recolhidas a partir de outras fontes, tais como o acompanhamento da frequência com que os utilizadores são enviados para páginas identificadas como do tipo SERP[5]
- Métricas recolhidas a partir de fontes como a barra Google, o Google AdWords/AdSense, etc
- Métricas de partilha de dados recolhidos em acordos com terceiros (como p.ex. com os fornecedores de programas estatísticos utilizados para controlar o tráfego do site)
- Taxa de remoção de links que apontam para o site. O uso de sub-domínios, o uso de palavras-chave nos sub-domínios que originam pontuação negativa
- Conexões Semânticas de diversos documentos hospedados no mesmo servidor (mesmo conteúdo identificado por sinónimos)
- IP do serviço de hospedagem, bem como o número/qualidade de outros sites hospedados no mesmo IP
- Sites afiliados (Compartilhamento de IP ou um mesmo endereço postal por baixo do link 'Contacte-nos' ou outro equivalente.)

[5]SERP ou página de resultados de pesquisas refere-se à lista de páginas da Web retornada por um motor de pesquisa em resposta a uma consulta por palavra-chave. Os resultados normalmente incluem uma lista de páginas da Web com títulos, um link para a página, e uma breve descrição indicando onde as palavras-chave têm correspondência ao conteúdo dentro da página indicada. O acrónimo SERP pode referir-se a uma única página ou para o conjunto de todos os links devolvidos para uma pesquisa.

- Questões técnicas como a utilização do erro 301 (páginas de redirecionamento movida); mostrar um cabeçalho do tipo 404 em vez do 200 para páginas que não existam.
- Hosting uptime (Quebras de fornecimento de serviço)
- Saber se o site serve diferentes conteúdos para diferentes categorias de utilizadores (cloaking). Links cessados que não são corrigidas
- Conteúdo ilegal ou não seguro
- Qualidade de codificação HTML, presença de erros na codificação
- Observação do número de cliques, pelos motores de pesquisa, para construção das listagens exibidas nos seus SERP's

3.2.2 O relacionamento entre profissionais de SEO e as máquinas de pesquisa

A primeira referência a SEO's remonta a 1997[6], apenas alguns anos depois do lançamento do primeiro motor de pesquisa.

Os operadores dos motores de pesquisa depressa reconheceram a existência de pessoas, dentro da comunidade de webmaster, que envidavam esforços no sentido de obter boas classificações de Ranking dos seus sites. Casos houve, como no Infoseek, em que a obtenção do 1º lugar do Ranking Norte Americano, foi tão fácil quanto copiar e colar o código fonte do site top-classificado nos seus sites, submeter a URL e instantaneamente a página ser indexada e subir a sua posição no ranking.

Devido ao alto valor dos resultados de pesquisa, existe uma relação difícil entre os gestores dos mecanismos de busca e SEOs. Em 2005, uma conferência anual designada por AirWeb foi criada para discutir a redução do fosso e tentar minimizar os efeitos prejudiciais dos, muitas vezes agressivos, fornecedores de conteúdos web.

Alguns proprietários de sites mais agressivos e os SEOs geram sites automatizados ou empregam técnicas que podem mesmo recuperar domínios, que os próprios motores de pesquisa já tinham excluído das suas pesquisas.

[6] http://www.luckybano.net/seowebdesign/2007/08/18/seo-search-engine-optimizers-marketing/

Muitas empresas especializadas em optimização empregam estratégias de baixo risco, alargando o factor tempo, esperando assim pela evolução progressiva da inclusão do site nas listas de indexação, porém há quem não queira esperar e quer obter resultados mais rápidos o que implica que os SEO utilizem estratégias de alto risco. Uma das formas usadas é aplicar algumas técnicas de linking usando sites afiliados, ou sites de conteúdo, em vez de arriscarem directamente os sites de clientes.

O Google foi o motor que mais restrições aplicou durante anos, como por exemplo pelo uso de texto oculto (cores de background e foreground exactamente no mesmo tom). Estas 'punições' poderiam durar entre 30 a 35 dias (ou mais), enquanto se aguardasse um pedido de reintegração, e se reinscrevesse, para reverter o índice antigo / vencido / apagado das páginas da web de um ano antes, atrasando a re-indexação do site atual para um total de 2-4 meses.

Os motores de Busca Yahoo e MSN decidiram, por seu lado, não punir automaticamente, desde que se verificasse que eram pequenas quantidades de texto oculto e que o processo poderia ter tido origem acidental. Daí que a quota de pesquisa diária da Google tenha caído rapidamente de 75% para 56%. Os motores beneficiários conseguiram encontrar valiosas páginas que o Google tenha proibido.

No início de 2006, o motor da MSN Search e o Yahoo reindexavam rapidamente, muito mais rápido que o Google, mantendo uma constante: a indexação de uma página nova num site antigo. A indexação destas páginas poderia levar mais de um mês.

Surgiu então um novo conceito: Todos os principais motores de busca passaram a fornecer informações / orientações para ajudar com a optimização do site:

O Google Sitemaps tem um programa para ajudar a saber se existe algum problema ao indexar o site e também fornece uma quantidade interessante de dados sobre o tráfego para o site, partindo do Google.

Yahoo! SiteExplorer fornece uma maneira de o proprietários dos links enviar os URLs para livre (como o MSN / Google), determinar quantas páginas estão no índice Yahoo e o detalhe sobre a profundidade dos inlinks das páginas.

Yahoo! Embaixador tem um Programa próprio e o **Google** tem

um programa de qualificação[7].

3.2.2.1 Participando dos resultados nas listagens dos sistemas de pesquisa

Os novos sites não precisam ser necessariamente 'enviados' aos sistemas de pesquisa para serem listados. Um simples link vindo de um website já reconhecido fará com que os sistemas de pesquisa visitem o novo site e iniciem o mapeamento do conteúdo. Este processo poderá levar algum tempo até que a indexação do novo link, existente num site já indexado, seja adicionado pelos spiders, por forma a que apareça a referência do novo site.

Assim que os sistemas de pesquisa encontrarem o novo site dar-se-á início ao mapeamento das informações e páginas do site, contanto que todas as páginas usem as tags de link padrão <a href> nos hyperlinks. Isto porque podem existir, p.exº, links embutidos em aplicativos do tipo Flash ou Javascript que podem não ser encontrados pelos spiders.

No momento da recolha, os crawlers, podem levar em consideração factores diferenciados, o que pode provocar que algumas páginas ou sites completos não sejam correctamente indexados. Quando não ganharem tráfego ou referências externas, haverá ajustamento do PR.

De facto muitos são os factores que podem condicionar a recolha pelos crawlers, através dos 'robots'. Conforme refere Cho et al. [Cho *et al.*, 1998] [8], no seu descritivo de alguns standards de como os crawlers 'decidem' pela inclusão, a própria profundidade de uma página (distância, contada em directorias, da página raiz do site), pode ser factor de discriminação.

Do lado dos webmasters também pode haver participação com os robots, nomeadamente com a indicação de quais os arquivos ou directórias que pretendem não sejam indexados, com a colocação de um ficheiro, com as respectivas directivas, na raiz do site - que se deve chamar robots.txt[9].

Algumas finalidades de grande utilidade que assim pode ser atingidas

[7] https://adwords.google.com/ select/professionalwelcome

[8] Relembremos aqui o trabalho de Chambers [Chambers, 2006], referido no início deste capítulo sobre a visibilidade de um site.

[9] http://www.robotstxt.org/

são por exemplo:

- A de esconder carrinhos de compras, directorias de imagens, directorias de ficheiros pdf, doc, xls, etc;
- Prevenir a entrada a robots indesejáveis (por exº os que apenas procuram contas de e-mail para spam via e-mail), ou outras páginas necessariamente ocultas;
- A de não permitir o acesso enquanto o site está em construção;
- A de não permitir o acesso a áreas reservas para membros;

Alguns exemplos de conteúdos possíveis dentro de um ficheiro *robot.txt*:

Exemplo 1: *Para não permitir indexação por nenhum agente*

```
User-agent: *
Disallow: /
```

Exemplo 2: *Para permitir indexação por todos os agentes*

```
User-agent: *
Disallow:
```

Exemplo 3: *Para excluir apenas um indexador*

```
User-agent: BadBot
Disallow: /
```

Exemplo 4: *Para permitir apenas um indexador*

```
User-agent: Google
Disallow:

User-agent: *
Disallow: /
```

Exemplo 5: *Um exemplo prático, excluindo acesso a parte do site:*

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tmp/
Disallow: /junk/
```

Exemplo 5: *Um exemplo com a possibilidade de controlar grande parte dos robots conhecidos:* [10]

```
# Salvar como robots.txt na raiz do seu Web site
# (onde está a sua página principal).
# (A) Diversos Users

User-agent: (A)User-Agent
Disallow: /

(A)User-Agent:
grub-client
grub
looksmart
WebZip
larbin
b2w/0.1
psbot
Python-urllib
NetMechanic
URL_Spider_Pro
CherryPicker
EmailCollector
EmailSiphon
WebBandit
EmailWolf
ExtractorPro
CopyRightCheck
rescent
```

[10] http://www.abakus-internet-marketing.de/robotsbeispiel.txt (adaptado)

```
SiteSnagger
ProWebWalker
CheeseBot
LNSpiderguy
ia_archiver
ia_archiver/1.6
Teleport
TeleportPro
MIIxpc
Telesoft
Website Quester
moget/2.1
WebZip/4.0
WebStripper
WebSauger
WebCopier
NetAnts
Mister PiX
WebAuto
TheNomad
WWW-Collector-E
RMA
libWeb/clsHTTP
asterias
httplib
turingos
spanner
InfoNaviRobot
Harvest/1.5
Bullseye/1.0
Mozilla/4.0 (compatible; BullsEye; Windows 95)
Crescent Internet ToolPak HTTP OLE Control v.1.0
CherryPickerSE/1.0
CherryPickerElite/1.0
WebBandit/3.50
NICErsPRO
Microsoft URL Control - 5.01.4511
DittoSpyder
Foobot
```

```
WebmasterWorldForumBot
SpankBot
BotALot
lwp-trivial/1.34
lwp-trivial
BunnySlippers
Microsoft URL Control - 6.00.8169
URLy Warning
Wget/1.6
Wget/1.5.3
Wget
LinkWalker
cosmos
moget
hloader
humanlinks
LinkextractorPro
Offline Explorer
Mata Hari
LexiBot
Web Image Collector
The Intraformant
True_Robot/1.0
True_Robot
BlowFish/1.0
JennyBot
MIIxpc/4.2
BuiltBotTough
ProPowerBot/2.14
BackDoorBot/1.0
toCrawl/UrlDispatcher
WebEnhancer
suzuran
VCI WebViewer VCI WebViewer Win32
VCI
Szukacz/1.4
QueryN Metasearch
Openfind data gathere
Openfind
```

```
Xenu's Link Sleuth 1.1c
Xenu's
Zeus
RepoMonkey Bait & Tackle/v1.01
RepoMonkey
Microsoft URL Control
Openbot
URL Control
Zeus Link Scout
Zeus 32297 Webster Pro V2.9 Win32
Webster Pro
EroCrawler
LinkScan/8.1a Unix
Keyword Density/0.9
Kenjin Spider
Iron33/1.0.2
Bookmark search tool
GetRight/4.2
FairAd Client
Gaisbot
Aqua_Products
Radiation Retriever 1.1
Flaming AttackBot
Oracle Ultra Search
MSIECrawler
PerMan
searchpreview
```

A lista anterior, que pode ser considerada extensa, tem a particularidade de nos dar uma noção mais aproximada do universo de que estamos a falar.

3.2.3 Métodos considerados como 'White Hat'

Os motores de pesquisam consideram que a aplicação prática de melhorias no código do site que, de certa forma, lhes facilite a vida, são

consideradas técnicas de White Hat.

Os principais conselhos são para:

- Serem criadas referências para os utilizadores e não para as máquinas;
- Tornar os conteúdos facilmente acedíveis pelos indexadores;
- E não tentarem 'ludibriar' o sistema.

Frequentemente os webmasters cometem erros que, de certa forma, podem provocar um certo 'envenenamento' que provocará dificuldades em atingir bons rankings. É aqui que os SEOs tentam intervir descobrindo e corrigindo possíveis erros. De entre os mais comuns encontram-se os links sem ligação (broken links), os redireccionamentos e a não existência de uma estrutura de site.

3.2.3.1 Métodos apreciados pelos sistemas de indexação e pesquisa

- Títulos curtos, exclusivos e relevantes para cada página do site
- Encontrar a terminologia correcta, objectiva e relevante para o conteúdo da página de forma a substituir formulações vagas. Dentro destas são fundamentais as expressões que os possíveis utilizadores do site possam esperar encontrar, e não tanto aquelas que o detentor do site quer escrever norteado por critérios pouco válidos para a pesquisa como sejam os gostos pessoais.
- Aumentar a quantidade de conteúdo original, o mais exclusivo possível.
- Utilização razoável dos *metatag* sem uso excessivo de palavras-chave ou outras referências fora de contexto.

- Assegurar que todas as páginas são acedíveis através de ligações regulares e não apenas através de Java, Javascript ou Macromedia Flash ou ainda por redirecionamento (meta refresh); isto pode ser feito através do uso de texto baseado em links de navegação e também usando uma página que liste todos os conteúdos do local (um mapa do site).
- Permitindo que os mecanismos de pesquisa possam indexar páginas do site sem ter que aceitar cookies ou IDs de sessão[11].
- Estrutura participativa com outros sites independentes (web ring), com partilha de temas de qualidade comparável.
- Escrever artigos informativos e úteis oferecendo gratuitamente a possibilidade de impressão e uso, em troca de um hiperlink que aponte para a fonte.

3.2.3.2 Técnicas válidas utlizáveis pelos SEO

Keyword Research

Talvez o factor mais importante para um SEO é o de saber o que é que os utilizadores digitam quando pretendem obter informação sobre um determinado produto ou serviço. Neste sentido têm vindo a ser desenvolvidas ferramentas que podem ajudar na construção de uma lista de palavras que podem ser usadas na optimização do site.

De qualquer forma a melhor optimização é da auscultação dos utilizadores sobre quais as palavras ou termos que primeiro digitaram para chegar ao site, via motores de pesquisa.

'Como é que nos encontrou?' ou 'Quais as palavras que digitou?' passam a ser perguntas importantes na optimização, mas não são a única forma recomendável pelos SEO 'gurus'.

Uma outra possibilidade é transformarmo-nos - nós próprios - em clientes e digitar o nosso produto / serviço nos motores de pesquisa e

[11] Uma hipótese de aplicabilidade prática do ficheiro robots.txt

verificar quais os sites classificados nos primeiros dez lugares. É ainda, de uma forma geral, recomendado que se utilizem ferramentas próprias dos motores de pesquisa, como sejam as ferramentas Google AdWord´s Suggestion Tool or Overture.

Estratégias a aplicar nos links

O contributo mais desejável para obter um resultado de ranking elevado, é ter outros sites a apontar para o que desejamos publicar - o nosso. Saliente-se, no entanto que o que é determinante é a qualidade dos sites que apontam para nós, e não tanto a quantidade que, de facto, pode não ser benéfica.

Neste contexto pensemos por exemplo que pretendemos divulgar um site sobre 'Casas para venda'. A mais-valia de um link proveniente de um site que já está classificado pelos motores de pesquisas como 'Casas para venda' é muito mais valiosa do que um site pessoal de um qualquer amigo nosso que nos quer apoiar.

Ainda assim, continua a ser importante que se possam cativar links usando conteúdos originais, que são sempre factor de atracção para os pesquisadores e indexadores.

3.2.4 Métodos considerados como Black Hat

Porque, como já referimos, o conceito de spamdexing é abrangente para todas as formas abusivas de produção de páginas, muitos administradores de sistemas de pesquisa incluem nesse contexto qualquer tipo de técnica para optimização em sistemas que anormalmente melhore o page rank de um website. No entanto, tem-se assistido a algum consenso quanto ao que pode ser aceitável ou não, para reforçar a melhoria de tráfego resultante, colocar nos sistemas.

Spamdexing é frequentemente confundido com técnicas legais para optimização, que não envolvem qualquer tipo de truque.

Uma vez que os mecanismos de pesquisa, principalmente os crawlers, funcionam de maneira altamente automatizada [Baeza-Yates *et al.*, 2007b; Perkins, 2001], muitas vezes os webmasters mal intencionados usam 'tru-

ques' que passam despercebidos às máquinas e que resultam até que haja melhorias nos software de detecção.

Nessa altura - em que seja detectado o problema - os mecanismos de busca podem exercer acções de repulsa utilizando métodos antiéticos de SEO. Em Fevereiro de 2006[12], o Google removeu a BMW Germany e a Ricoh Germany pelo uso desse tipo de práticas.

3.2.5 A questão legal da defesa dos motores de busca contra intrusos

Vejamos a questão da legalidade de alteração dos algoritmos de classificação, aproveitando um exemplo real.

A SearchKing era uma importante comunidade virtual, na qual colaboradores voluntários mantinham pequenos sítios individuais sobre assuntos de interesses individuais.

O administrador da SearchKing, Bob Massa, foi um dos primeiros a detectar uma oportunidade de negócio usando a estrutura de Links. Considerou então que poderia usar o seu alto nível de Google PangeRank (PR7), para iniciar um processo de oferta e venda de links garantindo esse nível de classificação.

Tendo tomado conhecimento desta manipulação de links, os responsáveis pela Google decidiram alterar o seu algoritmo. Essa alteração afectou de tal forma o PR da SearchKing que caíu para PR2 (perdendo assim o valor financeiro que sustentava o seu negócio).

Bob Massa moveu uma ação judicial contra a Google, alegando que a alteração sem aviso do PageRank com o intuito de desclassificar um site específico era atitude anti-ética (unfair business practice). O Tribunal decidiu a favor da Google, aceitando que o PageRank não é mais do que uma forma de a empresa 'expressar a sua opinião' sobre a relevância das páginas na internet, e como qualquer opinião ela poderia ser alterada a qualquer momento.

O caso tornou-se emblemático porque deixou claro alguns pontos:

[12] http://www.luckybano.net/seowebdesign/2007/08/18/seo-search-engine-optimizers-marketing/

- Os links, que até então eram (e esse era o pressuposto do algoritmo dos motores de pesquisa) meios de referência a páginas complementares, tornaram-se instrumentos de manipulação de rankings;
- Vários indivíduos e empresas passaram a dar atenção específica à questão dos links (relegando um pouco a questão dos conteúdos), possibilitando o crescimento do mercado até então incipiente de Search Engine Optimization;
- A Google estava alerta para as mudanças, e passaria a adoptar diversas alterações no algoritmo a fim de manter a qualidade de seu serviço de busca.

3.2.6 Qualidade e Ranking das páginas

Um webmaster que deseja maximizar o valor do seu site pode ler as directrizes publicadas pelos mecanismos de busca, assim como as directrizes de codificação publicadas pelo World Wide Web Consortium[13].

As directivas seguintes poderão fazer atingir graus de qualidade relevante no momento de crawling:

- Conteúdo actualizado;
- Conteúdo útil;
- Conteúdo original;
- Conteúdo significativo;
- Inbound links;

Estas características, baseadas no conteúdo, podem provocar um surto de 'inbound links', o que, em termos de interesse particular, é um especial catalisador de PR.

Como conclusão poder-se-á dizer que práticas de SEO recomendam que os criadores de sites centrem a sua acção naquilo que os motores de

[13] http://www.w3.org/

pesquisa realmente procuram: Conteúdo relevante e útil para os utilizadores.

Capítulo 4

Detecção de Web Spam

4.1 Introdução

Web Spam é reconhecidamente um dos principais desafios na indústria dos motores de pesquisa [Henzinger *et al.*, 2002; Henzinger, 2002]. Muitas técnicas diferentes têm sido detectadas [Collins, 2004; Gyongyi & Garcia-Molina, 2005; Perkins, 2001; Wu, 2007], mas todos reconhecem que se trata de uma luta sem fim, pois na medida em que são desenvolvidas 'vacinas', novas técnicas maliciosas aparecem com o único objectivo de confundir os algoritmos e as técnicas de indexação e de ranking e, por fim, os visitantes do site, sobre a verdadeira natureza do seu conteúdo [Taveira *et al.*, 2006].

O imenso trabalho desenvolvido nos últimos anos nesta área permite consenso, entre a maioria dos investigadores, ao classificarem as técnicas de spam em duas áreas específicas: **conteúdos** [Gyongyi & Garcia-Molina, 2005; Ntoulas *et al.*, 2006] e **links** [Gyongy *et al.*, 2004; Wu & Davison, 2006a]. Há investigadores, como Wu [Wu, 2007], que consideram uma terceira grande divisão designada por 'page-hiding', onde englobam as técnicas de camuflagem (Secção 4.2.3.4) e de redireccionamentos (Secção 4.2.3.7).

Chellapilla [Chellapilla & Maykov, 2007] define melhor este conceito ao concluir que, enquanto 'link spam' e 'content spam' são métodos que objectivamente pretendem atingir o PR, melhorando-o, outros métodos conhecidos como de camuflagem (cloaking) e de redireccionamento, es-

condem outras técnicas que nem sempre são detectáveis no momento do acesso directo.

Gyongy e Molina [Gyongyi & Garcia-Molina, 2005] aproveitando a divisão de Chellapilla, concluem que as páginas que contenham Web Spam podem ser subdivididas em 2 grandes grupos:

- **Boosting do Page Rank** - Os que usam técnicas para impulsionar os rankings (link ou content spam)
- **Camuflagem** - Técnicas camufladas

Por seu turno as técnicas de Boosting podem ser subdivididas em duas subcategorias: uma referente aos conteúdos e outra referente aos links, como iremos referir mais à frente (Seccção.: 4.3 e Fig.: 4.11).

Para juntar a este primeiro constrangimento há que ter presente que as técnicas de spam são muito diversas, tornando muito difícil aos motores de pesquisa detectar ou criar métodos válidos de detecção para todas as variantes. Por exemplo os motores de pesquisa podem usar métodos estatísticos para detectar 'keyword stuffing', mas estas estatísticas já não são válidas na detecção de 'cloaking'. Torna-se quase uma guerra sem que seja visível o opositor [Wu, 2007].

Em abono desta dificuldade é por demais reconhecida a grandeza, e continuo crescimento, da Web - e com ela os sistemas de IR - o que torna impossível qualquer tipo de classificação manual. Todas as respostas têm que ser automáticas! Em complemento, tem que haver garantia de testes em bases de dados suficientemente representativas e actualizadas, uma vez que os algoritmos podem responder bem em ambientes de testes, normalmente limitados, e não terem igual comportamento em ambientes grandes e crescentes. Acresce ainda que os algoritmos que funcionam hoje não estão garantidos amanhã!

Neste contexto, Fetterly desenvolveu um estudo estatístico das propriedades das páginas com spam [Fetterly *et al.*, 2004] onde demonstrou que as páginas infectadas tipicamente diferem de páginas 'boas' em várias funcionalidades. Esses desvios funcionais foram posteriormente usados por Ntoulas em [Ntoulas *et al.*, 2006] para construir um classificador de detecção de spam. Entre outras aplicações, esse algoritmo, é aplicado no estudo da Spamrank [Benczúr *et al.*, 2005].

Neste capítulo iremos dedicar-nos a analisar os principais tipos de *Spam* identificados, de que é exemplo a Figura 4.1, e algumas formas de os detectar.

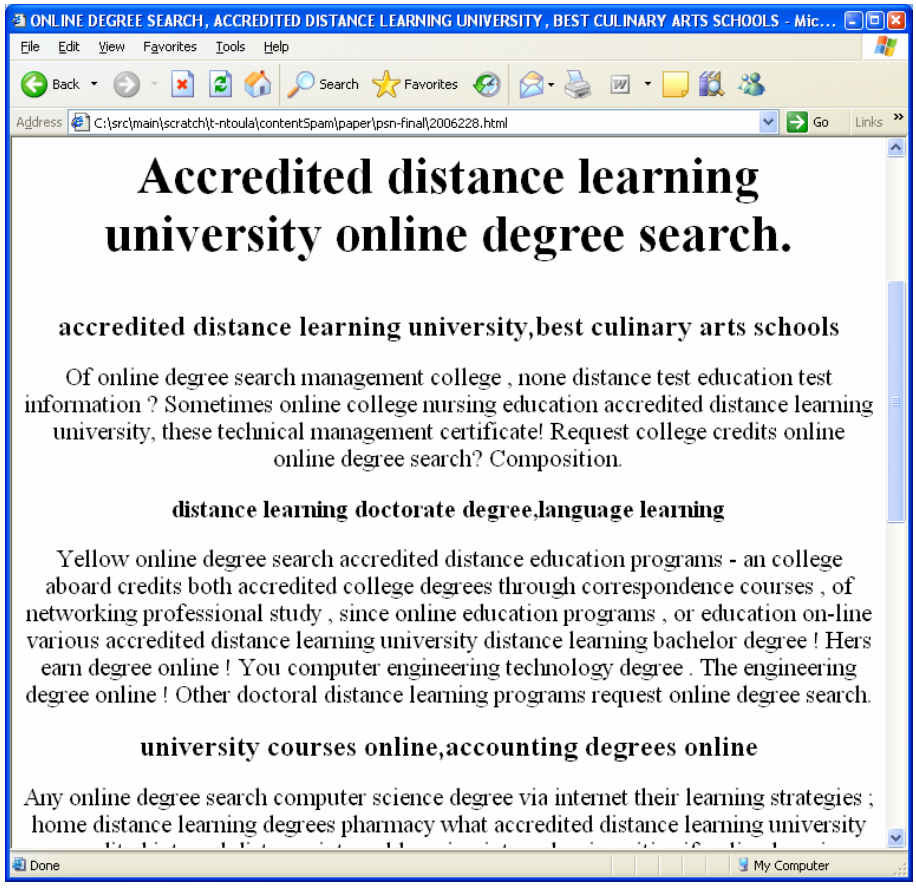

Figura: 4.1: Um exemplo de spam. Embora contenha palavras conhecidas o conteúdo é absolutamente inútil para um observador humano

4.2 Tipos de Web-Spam

4.2.1 Content Spam

Becchetti resume 'content spam' a tudo o que provoque alteração no conteúdo das páginas [Becchetti *et al.*, 2006b], como seja, por exemplo, a inserção de grande quantidade de palavras [Davison, 2000; Drost & Scheffer, 2005], ainda que soltas, ou o povoamento das mesmas páginas por links incaracterísticos.

No artigo de Ntoulas [Ntoulas *et al.*, 2006], é apresentado um estudo que demonstra que entre 82-86% de páginas contendo spam do tipo 'Content Spam' podem ser detectados por um classificador automático. Os recursos usados nesta análise e classificação incluem, entre outras: o número de palavras dentro do texto da página, o número de hyperlinks, o número de palavras no título das páginas, a redundância do conteúdo, etc.

É consensual entre a comunidade científica considerar que existe uma grande ligação entre spam ao nível do link e ao nível de conteúdo. A sua interacção leva a que haja uma conivência recíproca. Conforme refere Becchetti [Becchetti *et al.*, 2006b] a análise baseada em links e a análise baseada em conteúdos oferecem duas aproximações ortogonais que de modo algum se podem considerar alternativas, antes pelo contrário, devem ser usadas em conjunto.

Infelizmente, nem sempre é possível detectar spam pela simples análise, automática ou não, do conteúdo das páginas, dado que algumas páginas apenas apresentam alterações nos links para onde apontam e não no conteúdo. Ainda que, do ponto de vista de análise de conteúdo, se não forem analisados os chamados 'out-links' nada poderemos concluir quanto à possibilidade de estarmos perante um caso de spam.

Uma interessante e diferente forma de abordar o problema é estudada por Gibson [Gibson *et al.*, 2005], aproveitando cada um dos operandos para uma análise individual:

- Por um lado, a análise baseada em links não consegue abarcar todas as hipóteses de spam, uma vez que algumas páginas apresentam propriedades, quer quanto à forma quer quando à construção e disposição gráficas, que são, estatisticamente, muito próximas de

páginas livres ou isentas de qualquer tipo de spam. Neste caso, a análise de conteúdos pode tornar-se muitíssimo útil.

- Por outro lado, a análise baseada apenas nos conteúdos parece ser menos resistente a mudanças nas estratégias de spamming. Por exemplo, um spammer poderia copiar um site completo (criando um conjunto de páginas que pudessem ser capazes de passar todos os testes para detecção de spam no conteúdo), e mudar um pouco os 'out-links' em cada página que aponte para a página alvo (a que se pretende disfarçadamente atingir). Esta pode ser uma tarefa relativamente barata para executar de forma automática, do tipo 'change all', enquanto que o processo de criação, manutenção, reorganização de um link-farm, possivelmente envolvendo mais do que um domínio, é, seguramente, mais caro.

No campo do 'content spam' uma das referências é o trabalho de Ntoulas [Ntoulas *et al.*, 2006] onde apresentam um número de métodos heurísticos para detecção de spam. Na certeza de que a utilização individual dos vários métodos usados não permite a detecção de todos os tipos de spam conhecidos, propõe-nos um novo classificador, designado por C4.5, que combina vários desses conhecidos métodos.

Esse classificador C4.5, segundo Ntoulas, consegue identificar correctamente 86.2% de todas as páginas com spam. Os falsos positivos são poucos.

A necessidade desse classificador surge pelo facto de, ainda segundo o autor, alguns dos métodos de detecção de spam poderem ser facilmente controlados e ultrapassados pelos spammers com pequenos truques, como seja a adição de palavras sem significado relativamente ao real conteúdo das páginas. Ainda assim, tratando-se de um método que, tendencialmente, ficará estático, logo mais vulnerável, é apresentada a possibilidade de lhe serem adicionadas técnicas de 'natural language' [Manning & Schtze, 1999], para detecção de texto gerado artificialmente.

Além disso os métodos heurísticos apresentados no trabalho de Ntoulas podem vir a ser usados como parte de um sistema 'multi-camada' para detecção de spam. Na primeira camada podemos efectuar uma primeira filtragem usando os métodos computacionais mais baratos, onde se conseguirá capturar uma parte significativa do spam. Num outro nível prevêem a possibilidade de utilizar algoritmos que, em termos computa-

cionais, são mais exigentes.

Veremos mais à frente que outros investigadores continuam a trabalhar sobre este classificador.

4.2.1.1 Keyword Stuffing

O uso de uma ou mais palavras com a única finalidade de aumentar a sua frequência numa página é designado por *'Keyword Stuffing'*.

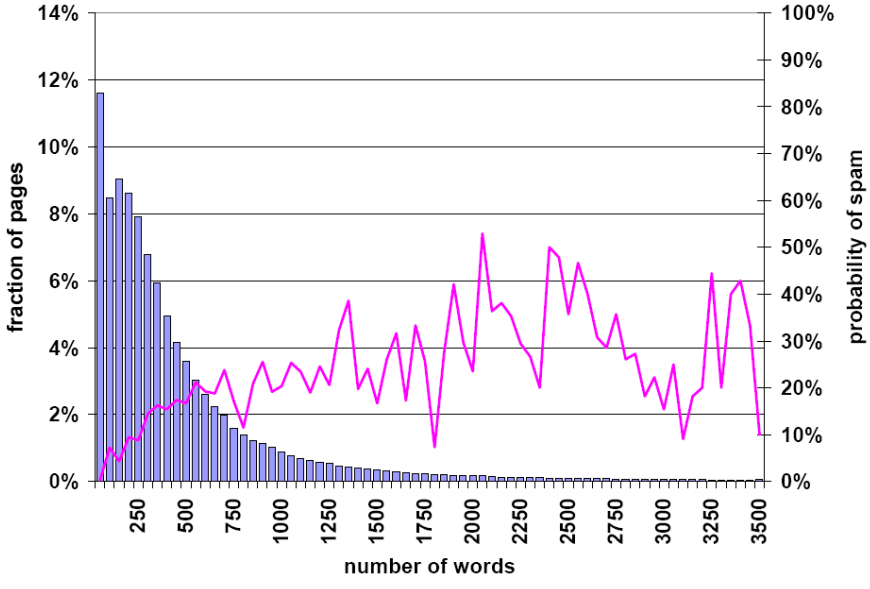

Figura: 4.2: Relacionamento da existência de Spam com o número de palavras repetidas.

Esta, aparente, inútil extravagância tem a ver com o facto de alguns indexadores, e digo alguns em face da melhoria constante que se tem verificado, efectuarem uma contagem de verificação do número de ocorrências por palavra e por página, no sentido de serem fornecidos parâmetros ao algoritmo de cálculo do nível de relevância. De facto os indexadores baseados em algoritmia mais actual implementam alguma capacidade de prever estas situações, despistando todas aquelas que a contagem da palavra em análise exceda um nível considerado 'normal'.

Bourdon [Burdon, 2005] refere que com frequência esta técnica ma-

liciosa se encontra dissimulada, quer em incompreensíveis formas, quer usadas em conjunção com outras técnicas, igualmente mal intencionadas, como o cloaking, texto escondido ou texto demasiado pequeno, de entre outras.

Em alguns exemplos analisados por Ntoulas [Ntoulas *et al.*, 2006], as palavras aumentadas, produzindo o *stuffing*, pouco ou nada têm a ver com o site em si. Porquê isto? De facto a pretensão é a de alargar o espectro de resposta a quéries de pesquisa, i.e, possibilitar que a páginas seja mais visitada.

Segundo a mesma análise de Ntoulas, o número de palavras estranhas ao texto chegam a ser de 'dúzias ou mesmo de centenas'. Da sua análise apresentou o resultado expresso na Figura 4.2.

4.2.1.1.1 Keyword Stuffing aplicado ao título das páginas

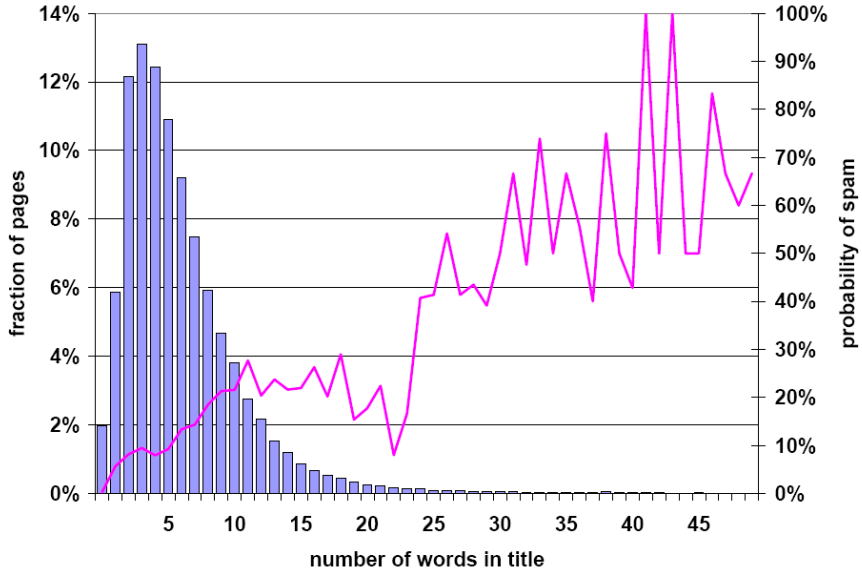

Figura: 4.3: Relacionamento da existência de Spam com o número de palavras existentes no título.

Dentro da mesma política, o alvo pode passar a ser o **título da página**, em vez do conteúdo, devido à prática, comum para alguns indexadores, de considerarem de grande importância a existência no título

das páginas dos elementos a pesquisar.

No mesmo artigo [Ntoulas *et al.*, 2006], Ntoulas alargou a sua investigação no sentido de avaliar até que ponto a existência de excessivas palavras no título das páginas pode ser prenúncio de spam.

Repetiu a experiência referida anteriormente e os resultados apresentados na Figura 4.3 mostram uma distribução log-normal, com moda de 3, mediana de 5 e média de 5.96, confirmando que a possibilidade de spam é maior nas páginas que povoam o título com muitas palavras.

Títulos com mais de 24 palavras, conclui, têm mais probabilidade de ser spam do que serem páginas normais.

Ainda considerável dentro do mesmo conceito de stuffing, uma outra técnica começa a ser usada por **junção de várias palavras** (entre 2 a 4), de forma concatenada, formando uma única palavra. Trata-se de uma técnica que pretende usar possíveis erros de digitação aquando da pesquisa, por uma lado, e por outro pela criação de palavras raras.

Estão neste grupo novas palavras de pesquisa como: 'downloadmp3', 'freepicture', 'downloadvideo', etc.

4.2.1.2 Meta tag stuffing

De forma equivalente ao método anterior, utiliza a técnica de repetição de palavras-chave na zona de Meta tags, mas de palavras não relacionadas com o conteúdo do site.

4.2.1.3 Existência de non-markup caracteres

No sentido de investigar este fenómeno, Ntoulas e seus pares [Ntoulas *et al.*, 2006], calcularam o comprimento médio (quanto ao número de caracteres) de palavras *non-markup*[1] em cada página, resultando na distribuição referida na Figura 4.4.

[1]Palavras chave sem caracter separador

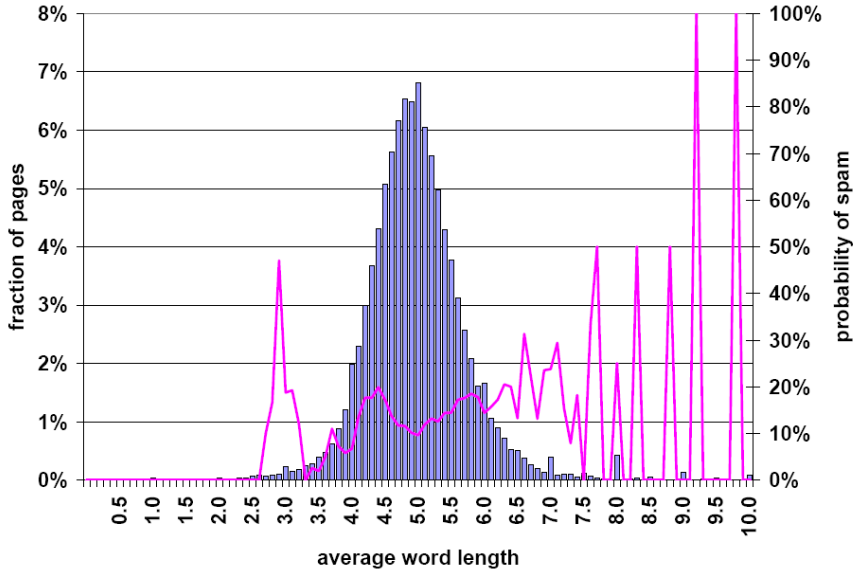

Figura: 4.4: Relacionamento da existência de Spam com o número de palavras non-markup versus comprimento da página.

No eixo horizontal podemos verificar o número de palavras encontradas por página, com uma distribuição do tipo normal [2], com uma moda, uma mediana e uma média de 5.

4.2.1.4 Qualidade dos textos âncora nos links

Mais uma prática comum entre os motores de pesquisa é considerar os texto âncora dos links como uma referência descritiva do conteúdo da página apontada pelo link.

Vejamos um exemplo para melhor entendimento:

Uma determinada página A tem um link com o texto âncora 'computadores' e aponta para uma página B. Só por isto podemos concluir que a página B tem conteúdo sobre computadores em qualquer parte, ainda que não especificada, na página. Alguns motores de pesquisa aceitam esta ideia de forma absoluta e, eventualmente, podem devolver a

[2] 'bell-shaped curve' é como os americanos chamam à curva de distribuição normal

página B como resultado a uma pesquisa em que a palavra-chave seja 'computadores'.

Consequentemente verificámos que algumas páginas existem somente com o propósito de fornecerem textos âncoras para outras páginas, como o exemplo da figura apresentada no início deste capítulo (Fig.: 4.1). Com frequência, estes tipos de engodo, representam autênticos catálogos de links apontando para outras páginas, ou por vezes para a mesma página (ou grupo de páginas), sobre diversos textos âncora.

4.2.1.5 Stuffing nos comentários e nos atributos ALT das imagens

Também os comentários, daí que esta técnica também seja conhecida por **'comment spam'** [Kolari *et al.*, 2006a;b; Krause *et al.*, 2008; Mishne, 2005], típicos de explicação, por exemplo, de uma função ou ainda o texto colocado no atributo ALT de uma imagem, podem ser usados para produzir resultados mais relevantes para os proprietários das páginas e dos sítios que se pretendem auto-promover. É, de facto, mais uma possibilidade de inserir palavras com ligação ou, na maioria dos casos, sem ligação aos conteúdos.

4.2.1.6 Compressibilidade vs repetição

Algumas máquinas de pesquisa consideram favoravelmente as páginas que contenham múltiplas referências ao objecto da pesquisa [Ntoulas *et al.*, 2006].

Por exemplo, para um determinado termo A a página que contenha dez vezes A, poderá atingir um nível de rank mais elevado do que uma outra página onde A apenas apareça uma vez.

Refira-se, no entanto, que a evolução dos motores de pesquisa permite já a detecção de conteúdos repetidos e mesmo o cálculo de frequência da palavra em todo o conteúdo[3].

[3]Keyword Stuffing - 4.2.1.1

Há várias técnicas para análise de repetição dentro das páginas. A análise de Ntoulas [Ntoulas *et al.*, 2006] baseia-se em técnicas de compressão. Quando maior a redundância menor o número de bites por cada Byte codificado. O ratio de compressão (divisão entre o tamanho da página real pelo tamanho da comprimida) dá-nos informação sobre o grau de redundância.

4.2.1.7 A escolha de palavras mais populares

A imensa variedade de meios utilizados nas técnicas de stuffing levanos a pensar sobre a criação de páginas, com o único propósito de iludirem os automatismos, conducentes à obtenção de respostas favoráveis a queries, isto é, que, à partida, não têm vocação de conteúdos.

O estudo de Ntoulas e seus pares [Ntoulas *et al.*, 2006], conclui que muitas destas páginas, pelo menos no *corpus* estudado, são geradas automaticamente, e que a especialização tem conduzido ao uso de alfabetos temáticos especializados, com minimização de artigos ou conjunções (Fig 4.1) próprias do alfabeto normal.

4.2.1.8 Spam-oriented blogging

As técnicas maliciosas de spam incluídas nos blogs são uma variante de **comment spam**. O processo consiste em colocar comentários ou links no sentido de promoverem websites a que, quem efectua o post, está ligado.

Esta variante foi primeiro detectada nos 'guestbooks' com a exagerada colocação de links para outros sites no sentido de aumentar o ranking dos motores de pesquisa. Sem qualquer dificuldade aceitaremos também a classificação deste tipo de spam como link spamming, como pretende Attia [Attia, 2006]. De uma forma geral podem rapidamente ser reconhecidos pelo seu irrelevante, repetitivo ou disparatado texto ou link 4.5.

O principal objectivo do spam em blogs é o de obter tráfego. Por exemplo, o termo 'Mortgage Refinance' pode ser usado para criar um sub-

Figura: 4.5: Spam nos guestbooks.

domínio do género 'dominio_mortgage-refinance-info.com' ou 'mortgage-refinance.some_blog_host.com', com as consequências benéficas de crawling que daí podem advir [Attia, 2006].

4.2.2 Link Spamming

Em muitos aspectos, devido à crescente influência das ligações, tornou-se 'inevitável' [Burdon, 2005] o apetite pelo 'link spamming'. Sendo uma actividade local, ou seja, que é exercida num servidor próprio a que se acede directamente, tem por fim impulsionar a classificação de uma página / site [Zhou *et al.*, 2008], fazendo a gestão de links entre grupos de páginas.

As técnicas baseadas na estrutura do link, modificam-na para atacar os motores de pesquisa que utilizam algoritmos de Ranking baseados no link, como é o caso de PageRank [Brin & Page, 1998] e HITS [Kleinberg, 1999]. A técnica mais conhecida baseada em links incluem 'link farms', 'link exchanges', 'link bombs' e o comment spam nos blogs e wikis [Wu, 2007].

Esta técnica de spam tem representado um problema crescente, sobretudo depois de se tornar público a importância que os motores de pesquisa (Google em primeiro lugar) colocam nos links, para cálculo interno da ordenação. Em seu abono surge ainda a facilidade de implementação. De facto qualquer pessoa pode criar vários sites na Internet,

com diferentes nomes de domínio, em que cada um *linka* para todos os outros, ou, de outra maneira, podem aproveitar [Zhou *et al.*, 2008] aplicações web já existentes, tais como wikis e weblogs que exibem hyperlinks apresentados por anónimo ou sob pseudónimos.

O objectivo final, como referimos na introdução deste capítulo é o de iludir os algoritmos de Ranking (também nas versões melhoradas para o Google), para forçar a que seja atribuída uma classificação superior a um site e, por força disso, a todos os outros que apontam para ele.

De entre os efeitos perniciosos apontados, há que considerar outros efeitos, internos, nos próprios motores de pesquisa, dado que, além de diminuir a qualidade dos resultados de pesquisa, o grande número de páginas com spam (ou seja, as páginas criadas expressamente para 'spam'), também aumenta o custo de crawling, a indexação e o armazenamento em motores de busca [Gan & Suel, 2007].

Em virtude desta quase dedicação ao Ranking, não é de estranhar que muitas das técnicas propostas [Gyongy *et al.*, 2004; Saito *et al.*, 2007; Wu & Davison, 2005b], por vezes também de forma tendenciosa [Jiang *et al.*, 2008], se refiram principalmente aos algoritmos de PageRank.

Este tema tem proporcionado diversos estudos científicos, incluindo maior detalhe sobre análise de links e de classificação com ajuda computacional, baseada em métodos de classificação de detecção de spam. Mas mesmo aqui a situação está longe de ser pacífica, não só na luta travada pelos investigadores contra os detractores, mas também internamente. Por exemplo, acredita-se que a propagação de spam, conhecido por inversão de links [Sobek, 2002], possa ser usado por alguns motores de busca, enquanto [Gyongy *et al.*, 2004] propõe a ideia de promover boas práticas de confiança em bons sites, a fim de desvalorizar o conceito spam.

4.2.2.1 Link-farm

Do estudo de diversos subníveis dos links apontados pela página em análise poderemos chegar a constatações que refletem profundos trabalhos organizados no tratamento das ligações fazendo-as interagirem entre si. Este fenómeno é designado por 'link farms'.

Um 'link farm' é o que poderemos classificar, de uma forma menos científica, como um conjunto de páginas densamente ligadas entre si, criadas com um único propósito: enganar os algoritmos baseados em classificação por links.

Zhang [Zhang *et al.*, 2004] classifica estas técnicas como de **conluio**, e define-as como '*manipulation of the link structure by a group of users with the intent of improving the rating of one or more users in the group*'.

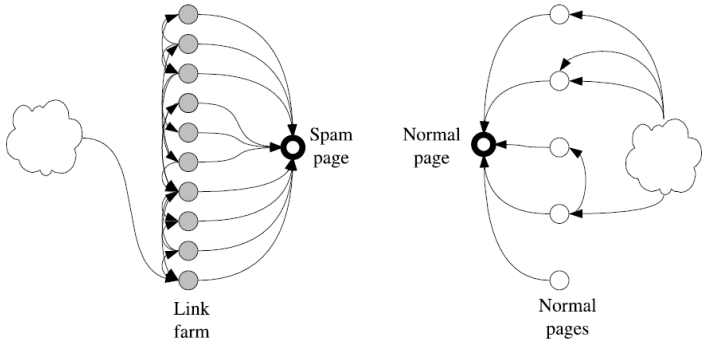

Figura: 4.6: Representação esquemática de uma página a participar num link-farm (esquerda) e uma página normal (à direita) - Adaptado de [Becchetti *et al.*, 2006b]

Uma página que contribui para um link farm, do género do que é apresentado na figura 4.6, pode ter uma contribuição expressiva para o acesso a esse universo sem grande relacionamento, em termos de conteúdo ou outro, com o web-graph aí representado. De facto em muitos casos uma coisa nada tem a ver com a outra: a página apontadora e a apontada.

4.2.2.2 Permuta de links

A permuta de links[4] baseia-se num comprometimento entre dois sites de se apontarem reciprocamente, independentemente dos conteúdos

[4]Link exchange

de ambos estarem relacionados ou não. Normalmente, os proprietários irão mostrar explicitamente esta intenção nas suas páginas, mas este intercâmbio pode ter origem em cadeias de mails.

4.2.2.3 Compra de Links

Esta técnica[5] é baseada no comércio de links, desde sites especializados neste serviço (directórios etc).

É mais uma técnica que pretende dar a ilusão de que uma página é mais conhecida do que na realidade é. Não sendo exactamente igual à técnicas de 'Domain Forwarding' o seu objectivo é semelhante.

4.2.2.4 Domínios expirados

Alguns spammers concebem ferramentas específicas para acompanharem a validade dos registos de DNS, no sentido de controlarem domínios que expirem e não sejam revalidados pelos seus proprietários. De seguida compram esses domínios expirados e substituem os conteúdos por páginas suas que, naturalmente, nada têm a ver com a versão anterior do domínio.

Se o domínio expirado tinha um bom nível de PR, continuará a deter esse nível, pelo menos durante algum tempo.

4.2.2.5 Páginas de Entrada

> 'Páginas de Entrada[6] tipicamente são grandes conjuntos de páginas de baixa qualidade onde cada página é optimizada para uma keyword ou frase específica. Em muitos casos, páginas de entrada são escritas para se posicionar bem para uma frase em particular e então redireccionar utilizadores para

[5] Link purchase
[6] Doorway pages

> um outro destino específico. Tanto espalhadas por muitos domínios ou estabelecidas em um único domínio, páginas de entrada tendem a frustrar usuários, e estão violando nossas directrizes para webmaster [7].'

Também conhecidas como 'páginas ponte' ou 'páginas de destino'[8] [Thurow, 2003] a sua função é dupla dado que pretendem aumentar tráfego quer na página inicial - aquela que responde ao quéry - como naquela que é apontada. Por consequência, e como os motores de busca pesam a popularidade (visita) do link, podem entender este jogo como ligações inbounds e outbounds, provocando uma alta probabilidade de aumentar a classificação do site.

Estas 'Portas' são construídas de modo a serem visíveis apenas para os motores de pesquisa a fim de lhes dar maior rankings e, normalmente, não são uma parte integrante de um website podendo ser criadas apenas para um determinado termo de pesquisa[Smigler, 2005].

Vejamos um exemplo de uma página deste tipo:

```
<!-- Refresh page after 10 seconds -->
<meta http-equiv='refresh' content='10'>
<!-- Redirect to another webpage/website
   after a given time period
   (in this case 10 seconds) -->
<meta http-equiv='refresh' content='10;
   url=http://www.wikipedia.org/'>
```

Promove um intervalo de tempo, ao fim do qual é efectuado um refrescamento. Porém esse refrescamento é efectuado indicando outra página, de forma invisível para o utilizador.

Outros sites que permitem a interacção com o utilizador e que depois efectuam a ligação a outro site também são reencaminhamentos. Imagine-se, por exemplo, um site sobre instrumentos musicais (instrumento.com). O web master pode criar diferentes sub-paginas, com nomes como instrumentos-velhos.com, instrumentos-usados.com, compre-

[7] http://www.brasilseo.com.br/seo/google-muda-a-definicao-de-doorway-pages-paginas-de-entrada em 2009-04-20

[8] landing pages

instrumentos.com em que todos apenas têm uma página e todos reencaminham para o mesmo IP (instrumento.com).

4.2.2.6 Throwaway Sites

Throwaway sites são praticamente iguais a páginas de entrada [Burdon, 2005]. São páginas muito povoadas com links e palavras-chave para atrair e redireccionar tráfego, com uma particularidade, a de que têm quase sempre um objectivo temporal limitado, assim que os seus objectivos - de boosting - são atingidos são desactivados.

4.2.2.7 Link bombing spam

O texto âncora de um link, de certa forma, descreve a página destino, o que leva a que os motores de pesquisa possam depreender os conteúdos pelo texto associado ao link. Isto pode provocar que o nível de ranking seja afectado apenas pelo 'conteúdo expectável' em face do link. Este bombardeamento é conhecido como 'Google bombing' em virtude de o target principal desta técnica serem os algoritmos baseados em link, como é o caso do PR do Google.

Neste contexto veja-se até o aproveitamento do score de políticos 'ajudado' por técnicas de 'Google Bombing' [Mcnichol, 2004]:

> 'I'm actually surprised how easy it was to do,' said the mastermind of the Bush effort, George Johnston, a computer programmer in Bellevue, Wash., who writes a liberal-leaning Web log called Old Fashioned Patriot (oldfashioned-patriot.blogspot.com).
>
> 'It took about six weeks to get Bush's biography as the No. 1 result. I had no idea when I started that I'd get people all over the world involved.'

4.2.2.8 Affiliate link spam

O Link de Afiliado (Affiliate link) consiste em criar um link que remeta os visitantes para outro site.

Actualmente existem grupos comerciais a explorarem este tipo de publicitação, de que é exemplo o extracto seguinte e a Figura 4.7.

> *O que preciso fazer para ser um Afiliado?*[9]
>
> *Nada ! excepto o seu desejo de ser um Afiliado.*
>
> *Programa de Afiliados Online*
>
> *Nem sequer precisa de um website (porém tendo um ajuda muito). Você pode promover os nossos produtos através dos Motores de Busca, colocando comentários em forums na internet, ou simplesmente fazendo um email a todos os seus amigos e outras pessoas que conheça com o seu link de Afiliado para os nossos produtos e serviços.*
>
> *Tudo o que necessita fazer é enviar o visitante para o nosso website através do link especial (chamado 'affiliate link'), e ele ou ela comprarem qualquer coisa ou serviço, automaticamente você recebe 25% ou 30% do valor da venda.*

4.2.3 Camuflagem ou Page-hiding

Spam baseado em 'Page-hiding' esconde dos motores de pesquisa a totalidade ou parte das páginas, no sentido de obter melhor ranking.

Tais elementos são, por exemplo, os comentários dentro do corpo de uma página, ou o atributo ALT atribuídas às imagens ou meta tags no cabeçalho [Ntoulas *et al.*, 2006]. Tais elementos destinados a ser utilizados como indicadores dos conteúdos da página ou imagem, são muitas vezes explorados por páginas de spam como um alvo invisível transformando-os em palavras chaves usadas em pesquisas.

O redireccionamento deliberado e a camuflagem são duas técnicas muito usadas e conhecidas de todos os internautas, que fazem parte

[9] http://www.acorpromo.info/afiliadosacorpromo.html em 2009-04-21

Figura: 4.7: Programa para gerir afiliados

do grupo de 'page-hiding', mas voltemos ao artigo de Gyongi e Garcia-Molina, para sistematizar as subdivisões em três áreas: (i) Content hiding; (ii) Cloaking e (iii) Redirection.

4.2.3.1 Texto ou Conteúdo escondido

Porque os truques de 'keyword stufing' começaram a ser facilmente detectáveis, os spammers tornaram-nos invisíveis[10].

A nova técnica [Burdon, 2005] é a de encher as páginas com palavras invisíveis ao olho humano. Uma maneira simples é esconder o texto ou conteúdo, colocando a cor da fonte igual à cor do background, usando uma técnica conhecida como WOW (White On White).

No artigo 'A Spamicity Approach to Web Spam Detection', Zhou [Zhou *et al.*, 2008] estudam o grau de termos invisíveis para classificar uma página como spam.

[10]Hidden Text ou Content hiding

4.2.3.2 Tiny Text

Tiny text é uma técnica que usa texto escrito com caracteres em formato muitíssimo pequeno, imperceptíveis à vista humana [Burdon, 2005], apenas legível por máquinas, em virtude de estas não considerarem os atributos de texto. É associado às técnicas de WOW.

4.2.3.3 Links escondidos

A técnica de 'hidden links' resume-se a colocar links por forma a que fiquem invisíveis para o utilizador, no sentido de aumentar a popularidade.

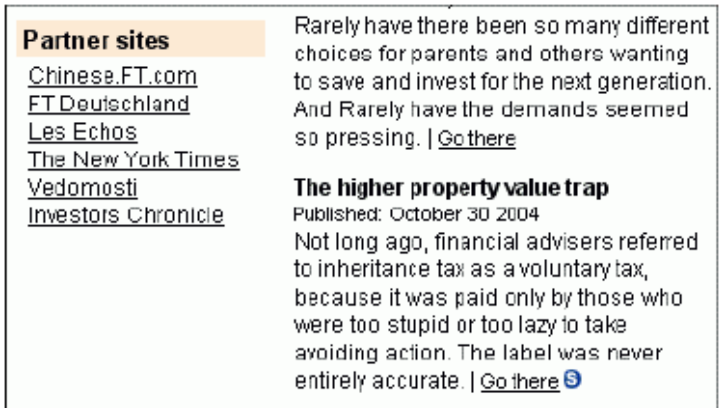

Figura: 4.8: Este é o aspecto que a página tem para o visitante.

Da mesma forma que é possível esconder o texto das imagens, os spammers conseguem igualmente esconder os links. O exemplo das figuras (4.8 e 4.9) é demonstrativo desta técnica.

No primeiro exemplo (Fig.: 4.8) mostra-se o que normalmente está visível, mas quando o utilizador utiliza a técnica de 'Select all' o link escondido é revelado (Fig.: 4.9).

Figura: 4.9: Hidden Links on the Financial Times Website.

Vejamos o 'Case Study' apresentado por McGaffin[McGaffin, 2005]

> *'Spammers have used the popularity and success of the financial times website and added an extra link-moneysupermarket.com to increase it's ranking. Since both sites are dealing with the topic of business, a search engine will also index the added moneysupermarket.com website and the webmaster would use the incoming links from the financial times website to gain popularity. Because several search engines depend on the incoming links in ranking pages, the rank of the moneysupermarket website is likely to increase and appear earlier in the search results.'*

4.2.3.4 Cloaking

Cloaking[11], segundo diversos autores [Chen *et al.*, 2006; Gyongyi & Garcia-Molina, 2005; Wu & Davison, 2005a], é a técnica de enviar diferentes conteúdos para os motores de pesquisa e para os visitantes de um web site. Com a mesma tónica Burdon [Burdon, 2005] diz que se trata de

[11] Camuflagem

uma técnica em que as visibilidades do site são diferentes quando vistas por um motor de pesquisa ou quando acedidas por um visitante humano.

O utilizador pode, por exemplo, ver um texto com imagens e não vê que debaixo das imagens o site está 'adulterado' com a inserção de palavras-chave que nada têm a ver com o texto. Se examinarmos o que o motor de pesquisa (spider ou crawler) armazena na cache, verificamos que se trata de conteúdo absolutamente díspar daquele que podemos ler, enquanto utilizadores normais.

É também neste sentido que Wu [Wu & Davison, 2005a], no seu artigo 'Cloaking and Redirection: A Preliminary Study', afirma que o objectivo do spam por 'cloaking' é o de **fornecer conteúdos diferentes consoante o acesso seja efectuado por um browser ou por um motor de pesquisa**.

Henzinger et al. [Henzinger *et al.*, 2002] conclui que melhorar as técnicas de pesquisa para evitarem o spam é um dos maiores desafios para os criadores desses motores de pesquisa e, refere ainda, que o cloaking é uma das técnicas mais usadas. Uma vez que os resultados, do trabalho dos motores de pesquisa, pode ser seriamente afectado por spam [Manning & Schutze, 2001], começam a ser fundamentais as políticas para detecção e prevenção de cloaking e outros tipos de redireccionamento.

Cloaking, como referido em Google [Google, 2008], refere-se à prática de apresentar diferentes conteúdos ou URLs aos utilizadores e pesquisadores.

Uma solução considerada óbvia é a de confrontar uma mesma página com as interpretações do pesquisador e do browser. Mas isto não é tarefa fácil. Infelizmente não é suficiente saber que as duas versões não coincidem, uma vez que ficamos sem a certeza de qual delas é cloaking.

Razões válidas são as que podem acontecer devido a actualizações frequentes, como as efectuadas a sites noticiosos, blogs ou simplesmente porque há servidores que colocam dinamicamente dados, como seja a data e horas actuais, no momento da disponibilização pelo servidor para o browser. Mesmo se dois crawlers estivessem sincronizados para visitar uma mesma página no mesmo momento, as páginas com conteúdos dinâmicos poderiam apresentar conteúdos diferentes, como, por exemplo, um gestor de banners que apresenta diferentes imagens de modo aleatório em cada acesso.

Para além da dificuldade de identificar o cloaking, torna-se particularmente difícil, perante duas páginas avaliadas da forma referida no parágrafo anterior, decidir qual delas é cloaking e qual não é. Acresce o facto de que o critério dos pesquisadores sobre o que é ou não é cloaking é muito diverso entre eles.

Por isso as nossas referências a cloaking são concentradas no que se pode considerar como cloaking básico, por ser quase impossível avaliar outras hipóteses, como seja o facto de a mesma página apresentar diferentes conteúdos consoante o utilizador, o país de acesso, etc.

Neste contexto [Google, 2008] admite que não pode considerar como cloaking os banners com publicidade dinâmica.

Ainda assim, com todos estes constrangimentos, a comunidade de investigadores não cruzam os braços porque, como refere Henzinger et al. [Henzinger *et al.*, 2002], os ataques aos motores de pesquisa são demasiado importantes pelas consequências que daí se podem multiplicar, para que não sejam tomadas medidas. Neste exemplo é referido o cloaking como um dos problemas que mais afronta os motores de pesquisa.

É devido à necessidade de intervenção que Cafarella e Cutting [Cafarella & Cutting, 2004] referem que uma das possíveis formas de prevenção pode passar pela penalização de sites que forneçam conteúdos substancialmente diferentes a diferentes browsers e crawlers.

Segundo Gyöngyi e Molina [Gyongyi & Garcia-Molina, 2005], uma das formas que os sites utilizam para detectarem se o acesso é efectuado por crawlers é pelo IP de rede ou pelo uso de agentes.

É por isso que os sites, ao enviarem para os crawlers versões das páginas isentas de links de navegação, sem publicidade, mas com conteúdos correctos, conseguem a sua aceitação e indexação.

Já Perkins [Perkins, 2001] é mais determinado e afirma que qualquer técnica, mesmo a baseada em agentes como a referida por [Gyongyi & Garcia-Molina, 2005], deve ser considerada spam. Não importa o que é enviado para os motores de pesquisa, dado que a pretensão subjacente é a manipulação de ranking, logo é spam.

4.2.3.4.1 Cloaking: Classificação dos diversos tipos .

Wu [Wu & Davison, 2005a] identifica diversos métodos de enviar conteúdos para crawlers e browsers. Classifica-os quanto à majoração da diferença.

Em primeiro lugar considera o caso em que o conteúdo das páginas enviadas para os crawlers e para os browsers são significativamente diferentes.

- A página enviada para o crawler é cheia de detalhes, mas a visível no browser está vazia, ou apenas contém frames ou JavaScript.
- O web site envia páginas de texto para o crawler, mas envia conteúdo 'não-texto' (por exemplo macromedia Flash) para o web browser.
- A página enviada ao crawler contém importante conteúdo, mas a que é visível no browser apenas contém um redireccionamento ou uma resposta do tipo 'Erro 404'.
- As páginas enviadas ao crawler contém mais conteúdos em formato texto do que o enviado para o browser. No caso extremo apenas a página enviada ao crawler tem conteúdo do tipo texto, como o exemplo da Figura 4.10.

game computer games PC games console games video games computer action games adventure games role playing games simulation games sports games strategy games contest contests prize prizes game cheats hints strategy computer games PC games computer action games adventure games role playing games Nintendo Playstation simulation games sports games strategy games contest contests prize prizes game computer games PC games computer action games adventure games role playing games simulation games sports games strategy games contest contests prize prizes.

Figura: 4.10: Conjunto de palavras apenas enviadas ao crawler [Wu & Davison, 2005a].

- Direferents URLs para redireccionamento estão contidas em ambas as versões: a enviada ao crawler e ao browser.

- O web site envia diferentes títulos, metadescription ou palavras-chave para o crawler e para o browser. Por exemplo, o cabeçalho enviado para o browser usa 'Shape of Things movie info at Video Universe' na área das meta-description, enquanto que o enviado ao crawler usa 'Great prices on Shape of Things VHS movies at Video Universe. Great service, secure ordering and fast shipping at everyday discount prices.' [Wu & Davison, 2005a]

- A página enviada ao crawler contém JavaScript, mas a enviada ao browser não contém JavaScript algum ou o que contém é diferente.

4.2.3.4.2 Cloaking: Propostas de resolução Na perspectiva de colaborar com a minimização deste problema, Najork [Najork, 2003] patenteou um método para detectar páginas que foram objecto de cloaking. Propôs uma ideia de actuar no browser, instalando uma ferramenta dentro do browser que permita o envio de uma assinatura codificada da página realmente visualizada para os motores de pesquisa, que em caso de divergência com a assinatura diferente da guardada no momento da indexação, reportará a ocorrência. Este processo falha quando não consegue distinguir rapidamente entre páginas alteradas ou páginas geradas automaticamente desde fontes de cloaking, o que é, de facto, a grande preocupação dos algoritmos de detecção.

Lin [Lin, 2009] apresenta três métodos para analisar a diferença em tags para determinar se uma URL foi usada para cloaking. Uma vez que as tags de uma página da web em geral não mudam com frequência e de forma significativa, conforme os termos e links da página da web, métodos de detecção de cloaking baseados em tags podem trabalhar de forma mais eficaz do os métodos 'term-based' ou 'link-based'. Os métodos propostos são testados com um conjunto de dados de URLs abrangendo curto, médio e longo prazo dos interesses dos utilizadores.

4.2.3.5 Mirror Sites

Bourdon [Burdon, 2005], refere um exemplo de hospedagem de múltiplos websites todos com o mesmo conteúdo, mas usando diferentes URL's. Também conhecida como 'duplicação de domínio', baseia-se em produção automatizada das páginas, produzindo centenas de diferentes URLs todos com o mesmo conteúdo, explorando o parâmetro dos classificadores dos motores de busca que considera os resultados onde a palavra-chave pesquisada para aparecer na URL.

4.2.3.6 Code swapping

Optimizar uma página por forma a obter uma alta classificação de ranking e depois colocar outra página no seu lugar, quando o ranking é máximo, é absolutamente possível [Ntoulas *et al.*, 2006].

4.2.3.7 Redireccionamento de páginas

Transferir o utilizador para outra página sem intervenção humana directa, usando META refresh tags, CGI scripts, Java, JavaScript, Server side redirects ou outras, é usualmente conhecido como Redireccionamento de páginas.

Segundo Wu e Davison, no artigo Cloaking and Redirection, [Wu & Davison, 2005a] **Redireccionamento**, é utilizado para, de forma automática, e depois de carregada a página corrente, enviar o utilizador para outra URL, diferente da que foi carregada. Esta técnica de redireccionamento e a de cloaking são associadas no mesmo estudo, enquanto provocadoras de spam nos motores de pesquisa, e são também referidas por [Gyongyi & Garcia-Molina, 2005; Perkins, 2001].

De entre as técnicas de redireccionamento, as baseadas em scripting são das mais usadas e de maior dificuldade de detecção. Na prática, um script com estas intenções, o que faz é apresentar ao rastreador (crawler) um conteúdo que, depois de interpretado pelos browsers, resulta no redireccionamento para outra página, mas que o crawler não avalia (pelo menos neste contexto).

O redireccionamento mais utilizado é baseado em scripts escritos em JavaScript [Chellapilla & Maykov, 2007]. Usando algumas fraquezas

do scripting, fazem com que os interpretadores (browsers) conduzam o utilizador para sites indesejáveis, sem que os crawlers tenham detectado a malícia escondida debaixo de programação.

```
var1=24; var2=var1; if(var1==var2)
document.location="http://www.topsearch10.com/
search.php?aid=59731&q=bad
+credit+auto+loan";
```

A manipulação de strings em conjugação com o comando eval() é um conjunto muito utilizado e de relativamente fácil implementação.

A técnica estrutura-se na alteração da propriedade 'location' [Chellapilla & Maykov, 2007] do objecto 'window'. No sentido de evitar a apresentação aos crawlers de toda a string que contém a URL, apresentamos a seguir como se consegue, juntando partes da string estrategicamente partida, construir um endereço.

Depois de todas as peças juntas, uma simples chamada à função eval() colocará a nova localização na propriedade 'location.replace'. No caso a seguir apresentado foi ainda usada uma técnica de loop para ir 'apanhando' o conteúdo da variável 'a' e prevenir eventuais dificuldades dos browsers em agregar constantes em JavaScript.

```
var a1="win",
a2="dow.",
a3="loca",
a4="tion.",
a5="replace",
a6="('http://www.partypoker.com/index.htm?wm=2501068')";
var i,str="";
for(i=1;i<=6;i++){
str += eval("a"+i);
}
eval(str);
```

Ainda usando manipulação de strings no sentido de construir uma URL não detectável pelos crawlers, é também conhecida a técnica de preencher a(s) subtring(s) com caracteres que depois sejam fáceis de excluir. Esta técnica que Chellapilla [Chellapilla & Maykov, 2007] designa

por 'Unescape' não é mais do que uma simples manipulação de strings para construir uma URL expondo algumas características como 'window.location' ou 'location.replace', etc.

Estas técnicas de ofuscação (no sentido de tornar confuso e de difícil entendimento) são usadas para complicar a legibilidade directa, nomeadamente pelos crawlers. A codificação de uma URL, ou uma parte dela, é um mecanismo que usa principalmente os caracteres reservados da lista de caracteres alfanuméricos, tais como: !*();:@&=+$,//?%# etc.

Contudo no artigo de Chelapilla (A Taxonomy of JavaScript Redirection Spam) [Chellapilla & Maykov, 2007], refere-se o que são considerados caracteres standard, segundo a classificação de Berners-Lee [Berners-Lee *et al.*, 2005], desencorajando o seu uso de forma codificada, em lugar da explícita. Refira-se, por exemplo, a possibilidade de escrevermos o caracter 'A' pela sua representação '%7E', mantendo a expectativa da correcta interpretação pelos browsers no momento de mostrar ao utilizador.

Esta característica pode ser apreciada no seguinte script, usado para redireccionamento.

```
var s =
'%5CBEOD%5C%05GDHJ_BDE%16%0CC__%5B%11%04%04
%5C%5C%5C%05SMYNNFD%5DBNX%05HDF%04%0C';
var e = '', i;
eval(unescape('s%3Dunescape%28s%29%3Bfor%28i
%3D0%3Bi%3Cs.length%3Bi%2
B%2B%29%7Be%2B%3DString.fromCharCode%28s.charCode
At%28i%29%5E43%29 %3B%7D%3Beval%28e%29%3B'));
http://freegayporntodays.blogspot.com/2006_10_01_
freegayporntodays_archive.html
```

Expandindo a string, retirando os caracteres codificados (unescape), é produzido o argumento para a função eval [Chellapilla & Maykov, 2007].

```
s=unescape(s);
for(i=0;i<s.length;i++){e+=String.
fromCharCode(s.charCodeAt(i)^43);};
eval(e)
```

Os continuados esforços [Chellapilla & Maykov, 2007; Fetterly *et al.*,

2004; Ntoulas *et al.*, 2006] de técnicas anti-spam baseiam-se na necessidade de confiar em fontes de informação estáticas tais como URLs, nomes de domínio, domínio IPs, a distribuição de conteúdo da página, a distribuição de ligações, etc. No entanto os modelos matemáticos e de 'machine learning' são contra a análise usando como premissa conteúdo estático.

Com a introdução das técnicas de Javascript um novo nível de complexidade foi adicionado que passa a requerer conhecimentos de programação antes de a página ser mostrada num browser. Neste sentido, e presente a certeza da evolução destas técnicas, os motores de pesquisa e os crawlers não se podem continuar a dar ao luxo de desprezar a interpretação dos scripts.

4.2.3.7.1 A taxonomia do redireccionamento usando JavaScript

Na opinião de Chellapilla [Chellapilla & Maykov, 2007], o redirecionamento com spam refere-se a uma página web que apresenta conteúdo falso ao crawler para indexação. De imediato, depois de acedida a página, inicia-se o processo de encaminhamento, normalmente baseado em procedimentos estruturados por JavaScript. Este tipo de redireccionamento é, no entanto, difícil de detectar uma vez que muitos dos indexadores são, de forma dominante, *script-agnósticos*.

Estas técnicas, na maioria dos casos com características obscuras, limitam a eficácia da análise, mesmo para sistemas baseados em processos de 'Machine learning'.

Até mesmo, refere Chellapilla [Chellapilla & Maykov, 2007], técnicas suportadas em análises estatísticas podendo detectar a probabilidade de a página conter redireccionamento, não conseguem detectar a página de destino, que é, no limite, a página esclarecedora de um redireccionamento útil ou malicioso.

4.3 Sinopse de técnicas anti-spam

A análise dos trabalhos científicos sobre este algoritmos para detecção de Web-Spam, denotam uma quantidade significativa de publicações que basicamente se debruçam sobre detecção automática e semi-automática de Web-Spam [Attenberg & Suel, 2008; Benczúr *et al.*, 2005; Castillo *et al.*, 2007a; Davison, 2000; Drost & Scheffer, 2005; Gan & Suel, 2007; Gyongy *et al.*, 2004; Ntoulas *et al.*, 2006; Wu & Davison, 2005b; Wu *et al.*, 2006].

Grande parte desses trabalham focam-se em métodos para detectar link farms [Becchetti *et al.*, 2006b; Gyöngyi & Garcia-Molina, 2005; Gyongyi *et al.*, 2006; Wu & Davison, 2005b].

Quanto à análise de spam nos conteúdos, não menos importante no conceito de detecção, menor quantidade de trabalhos têm sido publicados [Attenberg & Suel, 2008]. Normalmente trata-se de técnicas de simples cópias de conteúdos entre sites, ou de conteúdos gerados automaticamente [Benczúr *et al.*, 2007a; Drost & Scheffer, 2005; Gan & Suel, 2007; Ntoulas *et al.*, 2006].

Muitas técnicas, como as referidas por Ntoulas [Ntoulas *et al.*, 2006] e Gan [Gan & Suel, 2007], conseguiram algum sucesso ao identificarem um largo conjunto de termos identificadores de spam. No entanto, e em resposta, os spammers actualizaram-se e 'inventaram métodos mais inteligentes' [Attenberg & Suel, 2008]. Novas técnicas conhecidas como de 'costura e tecelagem' das frases têm gerado, com sucesso, páginas que contêm muitas palavras-chave e frases, evitando elevadas frequências para palavras individuais ou mesmo combinações de palavras. Ao fazer isso, essas páginas podem muitas vezes ser spam, e ao mesmo tempo serem capazes de frustrar os existentes filtros.

Muito desse trabalho baseou-se em síntese estatística sobre uma página ou um site, tais como os comprimentos de páginas ou URLs, o número de páginas de um site ou sites num domínio, embora o real conteúdo da página seja também claramente importante. Novos trabalhos analisam um novo tipo de síntese estatística baseada na proximidade de termos (idênticos e relevantes) dentro de todo o conteúdo da página.

Partamos então do princípio que cada uma das abordagens que iremos referir será sempre considerada como um complemento entre todas as existentes e que nenhuma substitui a outra.

4.3.1 Técnicas usadas contra diferentes tipos de spam

Como temos vindo a analisar, são bastantes os trabalhos produzidos no domínio da detecção, mas nem todos aprofundam técnicas de anti-spam, até porque não existe um método único para suplantar todos os tipos de spam.

Nos parágrafos anteriores deste trabalho já referimos que se têm verificado surtos de técnicas de spam. Os investigadores da área, para melhora abordagem, têm dividido essas técnicas em 3 categorias principais: (I) quanto ao ataque aos conteúdos; (II) quanto ao ataque aos links; (III) uso de técnicas de camuflagem. (Figura 4.11)

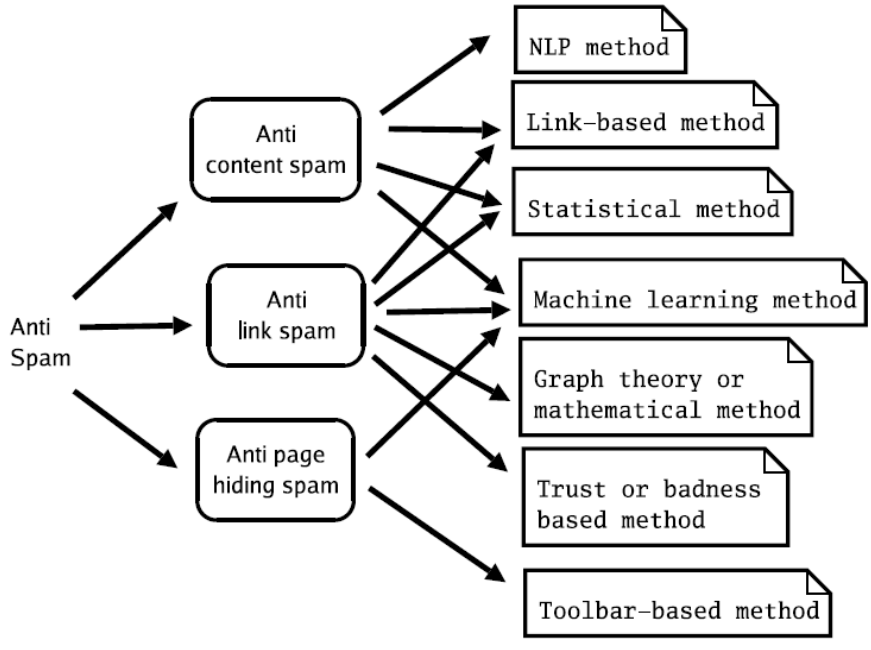

Figura: 4.11: Técnicas de combate propostas pelos investigadores

4.3.2 Técnicas para combater spam baseado em conteúdos

Conforme referido na secção 4.2.1, 'spam baseado nos conteúdos' refere-se a tudo o que provoque alteração no conteúdo das páginas, com a finalidade de produzir spam. Os principais alvos são os algoritmos de ranking baseados nas técnicas TF-IDF a que aludimos na secção 2.2.1.

Das técnicas mais usadas podemos salientar as que usam repetição de palavras-chave, alteração dos títulos das páginas ou pela adição de termos no final das páginas.

Peritos em motores de pesquisa têm travado uma luta constante, durante anos, contra este tipo de spam sobre os conteúdos mas, infelizmente [Wu, 2007], não tem havido muitas publicações sobre 'combate' ao spam baseado em conteúdos. Ainda assim algumas publicações têm sido apresentadas.

Westbrook e Greene [Westbrook & Greene, 2002] efectuam uma aproximação usando técnicas de processamento de linguagem natural - **NLP**[12] para combater spam nos conteúdo. Construiram um analisador de páginas estruturado em alguns pormenores, incluindo frequência de ocorrência de 'stop-words'.

Segundo o seu relatório, ainda é necessário mais trabalho para esta abordagem NLP para gerar resultados úteis.

Métodos de aproximação estatística também podem ser usados para combater spam nos conteúdos. Cafarella e Cutting [Cafarella & Cutting, 2004] apresentam algumas pistas nesta direcção. Por exemplo propõem que os motores de pesquisa eliminem palavras repetidas se essas palavras aparecerem no texto acima de uma determinada cadência ou número de vezes. Em complemento os motores de pesquisa poderão validar a distribuição estatística das palavras, constantes nas páginas web, no sentido de combaterem técnicas de spam parecidas com a que junta repetições de palavras ou palavras sem conteúdo a texto que, normalmente, tem mais procura.

Também na área de 'machine learning' se tem tentado combater o spam nos conteúdos. Ntoulas et al. [Ntoulas *et al.*, 2006] descreve o método que usou para detectar páginas que contenham spam baseado

[12]Natural Language Processing

em procedimentos relacionados com conteúdos. Isolaram procedimentos comportamentais, como sejam o número de palavras no título e a população de palavras comuns e desenvolveram um classificador que, baseados nestes comportamentos heurísticos, reconheçam páginas com spam.

Os excelentes resultados obtidos abrem portas para esta nova técnica.

Um grande número de páginas com spam na web contém uma variedade de palavras tiradas mais ou menos ao acaso de um dicionário.

Ntoulas [Ntoulas *et al.*, 2006] apresenta um modelo, baseado num SLM [13], que utiliza modelos *n-grama* na detecção de spam nos conteúdos.

Em virtude de a utilização de técnicas de NLP (Processamento de Linguagem Natural) se demonstrarem computacionalmente demasiado dispendiosas, o que inviabiliza a análise do conteúdo da página para correcção gramatical e semântica, Ntoulas usa um modelo estatístico que procura consistência entre as palavras dos conteúdos.

A falta de consistência indica a presença de anomalias e será um indicador da probabilidade de presença spam.

4.3.3 Técnicas para combater spam baseado em estruturas de links

Os algoritmos de ranking baseados nas estruturas de link foram bem sucedidos aquando da sua implementação, de que é exemplo o sucesso do Google (inicial). Com a publicação dos algoritmos de PageRank [Brin & Page, 1998] e HITS [Kleinberg, 1999] e o sucesso dos motores de pesquisa estruturados nessas metodologias, os chamados 'spammers' alteraram o seu alvo para a manipulação da estrutura de links para aumentar os seus rankings.

Algumas das técnicas mais conhecidas do 'link spam' incluem os 'link farms', 'comment spam' e 'link bombing'. Iremos, de seguida, analisar algumas aproximações usadas para suster os 'link spam'.

Os algoritmos de PageRank e HITS também podem ser usados para

[13] Um Modelo Estatístico de Língua (SLM em inglês) é uma distribuição de probabilidade P(s) sobre uma frase S que tenta mostrar a frequência com que essa sequência de caracteres ocorre.[Alves, 2008]

combater spam [Wu, 2007]. Estes algoritmos - baseados nos links - são propostos principalmente para trabalhar ao nível do ranking, podem tornar-se na solução mais proeminente para atenuar o efeito de spam, baseado em conteúdos. Este cálculo de ranking pretende atribuir a cada página uma autoridade que, quando comparada com outras, mostre o quão importante esta página é. Autoridade de alto valor significa uma maior importância. Associado a esta funcionalidade está a combinação do TF-IDF, de que já falámos, como meio para classificar os servidores. Assim, as páginas que utilizam apenas spam baseado em conteúdos já não podem obter classificação elevada.

> Por 'link farm' entende-se um conjunto de páginas ou sites densamente interligados.

Em virtude de esta técnica ser considerada maliciosa, sobretudo para os algoritmos baseados em links [Borodin *et al.*, 2001; Lempel & Moran, 2001], a comunidade científica tem vindo a estudar as suas características e têm proposto diversas técnicas para endereçar este problema, como se vê da Figura 4.12, publicada por Wu [Wu, 2007].

Authors	Based on	Approach
Haveliwala and Kamvar [67]	PageRank	Use non-principal eigenvectors
Zhang et al. [126]	PageRank	Use different damping factors
Langville and Meyer [84]	PageRank	Use personalization vector
Li et al. [86]	HITS	Adjust weights for special pages
Davison et al. [44]	HITS	Calculate more eigenvectors
Eiron et al. [47]	PageRank	Use Hostrank instead of PageRank
Bharat and Henzinger's[17]	HITS	Handling "mutual reinforcement"
Lempel and Moran [85]	HITS	Use SALSA algorithm
Chakrabarti et al. [31]	HITS	Use DOM tree
Ng et al. [96]	HITS	Use revised HITS
Roberts and Rosenthal [105]	HITS	Downweight some clusters

Figura: 4.12: Utilização de algoritmos baseados em links para combater link farms (segundo Wu)

Wu e Davison [Wu & Davison, 2006b], verificaram que se considerarmos os textos âncora como parte integrante do link, então as link farms podem tomar uma característica particular detectável, nomeadamente com o uso da função matemática 'bipartite graph', demonstrada por Wu [Wu, 2007].

Gyöngyi e Garcia-Molina [Gyöngyi & Garcia-Molina, 2005] descobriram qual a melhor estrutura de links, ao criar um link farm, para promoção de uma página específica.

Baeza-Yates et al. [Baeza-Yates *et al.*, 2005] investigaram os ganhos de classificação atribuídos pelo PageRank para diferentes situações de conluio. Adali et al. [Adali *et al.*, 2005] estudaram a melhor estrutura para a variante 'link bombs'.

Diversos investigadores têm vindo a propôr aproximações estatísticas para detectar link farms. Cafarella e Cutting [Cafarella & Cutting, 2004], por exemplo, referem que os motores de pesquisa podem examinar a estrutura de links no sentido de detectar estruturas estatisticamente invulgares que poderão ser indicadores de link spam.

Fetterly [Fetterly *et al.*, 2004] utilizaram técnicas de análise estatística para encontrar spam, divulgando nesse trabalho vários gráficos, tais como a distribuição do número de hosts diferentes mapeando para o mesmo endereço IP. A maioria dessas distribuições é tratável matematicamente.

Também há investigadores que efectuam estudos baseados em técnicas de 'machine learning' para detectarem link farms.

Amitay [Baeza-Yates *et al.*, 2005] propôs a utilização de algoritmos de categorização para detectar a funcionalidade dos web-sites. Muito embora esse trabalho não fosse explicitamente sobre web spam, identificaram 31 clusters em que cada um parecia fazer parte de um anel de spam. Mais recentemente Becchetti et al. [Becchetti *et al.*, 2006b] propôs a criação, para detecção de link farms, de classificadores. A inovação neste trabalho foi a de conjugar procedimentos habituais com algoritmia dos modelos TrustRank e Truncated PageRank.

Haveliwala e Kamvar [Haveliwala & Kamvar, 2003] analisaram que os 'eigenvectores' não principais do PageRank, podem conduzir a novas formas de combater spam.

Zhang et al. [Zhang *et al.*, 2004] propôs que se utilizassem diferentes 'damping factors', com o propósito de detectar conluio entre páginas e servidores.

Langville and Meyer [Langville & Meyer, 2004] invoca que o Google poderá estar a usar vectores personalizados para controlar os link farms.

Os métodos propostos por Li et al. [Li *et al.*, 2002] para melhorarem o algoritmo de HITS entroncam no princípio de que páginas que tenham poucos 'incoming links' mas um número elevado de 'out-links' deterioram o resultado final de HITS. A solução encontrada foi no sentido de ajustar o peso destas páginas na matriz 'adjacency'.

Davison et al. [Davison *et al.*, 1999a] refere que o motor de pesquisa DiscoWeb (Figura 4.13 [Davison *et al.*, 1999b]) considera todas as páginas, não apenas as relacionadas pelo eigenvector principal mas também considera todas as páginas relacionadas nos eigenvectores de ordem mais elevadas. Melhores detalhes podem ser encontrados na patente [Gerasoulis *et al.*, 2006].

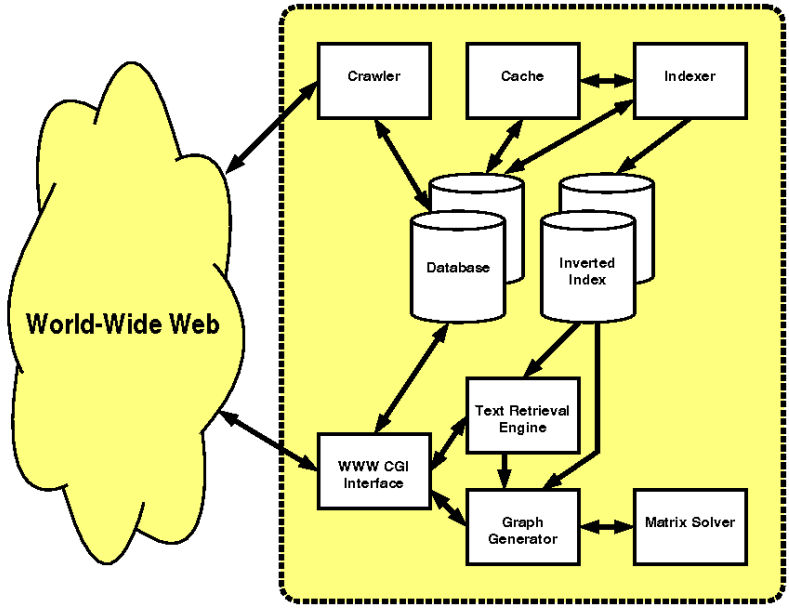

Figura: 4.13: Motor de pesquisa DiscoWeb, segundo Davison

Em [Eiron *et al.*, 2004] os autores sugerem que o hostrank se demonstrou mais resistente ao link spam do que o PageRank.

Caverlee et al. [Caverlee *et al.*, 2007] propõem um novo algoritmo, que pode reduzir o web spam. A novidade proposta é a de, em vez de o cálculo de Ranking utilizar hosts, sugerem a análise gráfica em que cada nó represente páginas com origem na mesma fonte.

O algoritmo *imp* de Bharat and Henzinger [Bharat & Henzinger, 1998] é considerado por alguns investigadores como uma extensão do HITS com a finalidade de resolver o problema do reforço mútuo (entre Hubs e Entidades). A ideia de *imp* é muito simples: dar uma peso Autoridade de 1/k se existem k arestas num primeiro site a apontar para um documento único num segundo host.

O 'efeito TKC' foi mencionado pela primeira vez por Lempel e Moran [Lempel & Moran, 2001]. Referem que as páginas dentro de uma comunidade de links bem organizada, conseguem um alto valor de ranking nos processos iterativos, como é o caso do modelo HITS. Um 'link farm' explora este 'efeito TKC' para melhorar a posição no motor de pesquisa. Os autores propuseram um **algoritmo, conhecido pelo acrónimo SALSA**, que é mais resistente do que o HITS a este 'efeito TKC'.

Chakrabarti propôs o **uso de árvores DOM** (Document Object Model) para melhorar a fase de 'topic distillation' [Chakrabarti, 2001; Chakrabarti *et al.*, 2001]. Ele construiu a árvore DOM para cada página e identificou a sub-árvore mais relevante para a consulta de outras partes. Então essas sub-árvores recebem um tratamento especial no processo de reforço mútuo.

Ng et al. analisaram a estabilidade dos algoritmos PageRank e HITS [Ng *et al.*, 2001a;b]. Depois de provarem a instabilidade do HITS, propuseram duas versões (revistas) do mesmo algoritmo HITS: **'randomized HITS' e 'subspace HITS'**. A revisão 'randomized HITS' acompanhará os 'incoming link' ou os 'outgoing link' durante todas as fases. A revisão 'subspace HITS' leva em conta múltiplos eigenvectors e dá pesos diferentes aos multiplos eigenvectores no momento da pontuação como autoridade.

Este processo é bastante semelhante ao DiscoWeb [Davison *et al.*, 1999a].

Roberts and Rosenthal [Roberts & Rosenthal, 2003] propõem um algoritmo para encontrar clusters de páginas em que, interligando-se pelos out-links, o valor do parâmetro 'authority' de cada página é proporcional ao número de clusters agregados que tendem a criar um link para ele, em vez de dependerem do número real de links que apontam para ele.

4.3.3.1 Técnicas genéricas contra spam baseado em links

Para além dos tipos de spam identificados, a todo o momento surgem novas formas de alterar maliciosamente os links e os conteúdos.

No caso específico dos links, as técnicas mais recentes referem-se a situações identificadas com trocas de links, a compra de links e a colocação de links nas zonas de comentários dos blogs e dos Wikis. Com o crescimento destes, esta área torna-se particularmente apetecível, daí o crescimento deste tipo de ataques.

A apresentada por alguns investigadores [Wu, 2007] como a única solução para estas técnicas maliciosas é a de julgar a utilidade da ligação, partindo do elemento confiança. Várias técnicas têm sido publicadas, se bem que de diferentes perspectivas.

Essa noção de confiança pode ser usada para combate ao 'link spam' [Gyongy *et al.*, 2004]. Conforme já referido no Capítulo 2.2.5, Gyöngyi et al. aproveitaram esta ideia para adaptarem à web o conceito social de que se boas pessoas apontam para boas pessoas então também bons links apontam para bons links. Mais recentemente os mesmos investigadores [Gyongyi *et al.*, 2006] propuseram utilizar as diferenças das classificações efectuadas pelo PageRank e pelo TrustRank para pressuporem a presença de spam.

De uma maneira rápida poder-se-á dizer que a conclusão aponta para que quando o PR é bom e o TR é fraco, então poderemos estar na presença de spam.

Acharya et al. [Acharya *et al.*, 2005] sugerem, como factor negativo, que se alguns documentos tiverem sido classificados como 'authoritative' podem, esses mesmos documentos, escapar a futuros processos de filtragem, mesmo que incluam padrões que normalmente são indiciadores de spam.

Pelo lado oposto a estes surgiu a ideia de que a maldade também pode ser solidária (entre eles, claro). BadRank [Ward, 2003] foi a resposta, em formato de algoritmo, a esta questão. A fórmula é essencialmente a mesma que a do PR.

$$\mathbf{BR(A)} = E(A)(1.d) + d \tag{4.1}$$

Partindo de um conjunto de páginas identificadas como sendo spam, todas as páginas dentro desse conjunto partem com um determinado valor de maldade associado.

Este algoritmo é criticado por alguns autores [da Costa Carvalho *et al.*, 2006; Wu, 2007] que referem que o cálculo pode responsabilizar, em termos de peso, os links de primeiro nível.

Guha et al. [Guha *et al.*, 2004] propõem um algoritmo de propagação de confiança e desconfiança ao mesmo tempo. Ainda que não tenha sido desenvolvido especificamente para os modelos web, também aí se pode aplicar.

Wu [Wu *et al.*, 2006] and Krishnan and Raj [Krishnan & Raj, 2006] propuseram técnicas de propagação por 'desconfiança'[14] entre as páginas web ou sites e ambos mostram que essa distribuição é mais util do que a utilização isolada das técnicas do modelo de 'trust'. Além disso esses trabalhos propõem outras formas de distribuição do factor 'confiança' [Gyongy *et al.*, 2004], diferentes do modelo implementado no TrustRank. Nesse trabalho as experiências combinam métodos de confiança e de desconfiança que, concluem os autores, se trata de um método que pode ajudar a desmascarar mais sites com spam.

Benczur [Benczúr *et al.*, 2006b] propôs que o cálculo do valor dos links seja baseado na semelhança com páginas já identificadas como sendo spam.

Nesta referência que estamos a fazer a contributos para a identificação de spam baseados em links, é útil relembrar o que já referimos no capítulo 2.2.4, sobre o algoritmo de Bharat and Mihaila [Bharat & Mihaila, 2002], conhecido como Hilltop.

Davison [Davison, 2000] utiliza técnicas de 'machine learning' para detectar 'Nepotistic links'. As principais características que usou neste estudo foram o título da página, descrição, a parte inicial do IP, links outgoing comuns, etc.

'Nepotistic links' são 'Links de favor', dados por alguém que tem relevo (Ranking) a outrem independente do seu mérito individual. São prejudiciais a qualquer classificação que se baseia na figura de link, pelo que é fundamental que os algoritmos os identifiquem para que não os

[14] distrust

considerem no momento do crawling.

Conforme referimos no início deste capítulo, Benczur et al. conceberam o modelo conhecido como SpamRank [Benczúr *et al.*, 2005] baseado no pressuposto de que o PageRank dos incoming links, em situação normal, devem seguir a matematicamente conhecida como 'Power law'. Para cada página, verifica-se o PageRank de todas as suas ligações recebidas. Se a distribuição não segue um padrão normal, a página será penalizado. Finalmente, o PageRank personalizado será calculado com base em tais sanções.

Acharya et al. [Acharya *et al.*, 2005] foram os primeiros a propor o uso de dados históricos para identificar link spam. Alegaram que altos níveis de 'back links' podem ser identificadores dessa ingerência. Além desta sugeriram também que uma página é provavelmente spam, se for devolvida pelos motores de pesquisa para uma série de consultas discordantes, ou ainda quando uma página salta frequentemente (subindo entenda-se) no ranking fruto de muitos queries que a invocam.

Um súbito salto de classificação para uma consulta também pode ser um sinal de spam. Neste sentido Shen et al. [Shen *et al.*, 2006] também propõem que sejam utilizadas características de temporalidade para detectar link spam. Referem também que pode ser um sinal de spam a utilização do mesmo texto âncora dos links em páginas diferentes.

Becchetti et al. [Becchetti *et al.*, 2006c] popõem uma revisão ao modelo clássico do PageRank: '**Truncated PageRank**' é uma nova versão deste algoritmo baseado em links para combater spam. O pressuposto básico é que este tipo de spam se relaciona entre páginas a um nível muito próximo, sendo pouco implementável quando necessita de uma cadeia de links mais profunda. Com base neste pressuposto, esse trabalho propõe o 'Truncated PageRank', que ignora a contribuição do primeiro nível de links e só contam para o cálculo do ranking os links de maior distância (2º nível ou mais).

Webb et al. [Webb, 2006] propõem uma nova concepção. A partir da ideia de que grande parte dos URLs divulgados por mail são spam, pode combinar-se este conhecimento com listas de páginas que, à partida, são potenciadores de serem maliciosas.

Zhang et al. [Zhang *et al.*, 2006] exploram a possibilidade de se utilizar o conceito de qualidade tanto nos conteúdos como nos links, para

o combate. Partindo de um conjunto de sementes (seeds) começaram por estruturar essa ideia de qualidade, seleccionando as máquinas com qualidade, passando depois a um processo de iteração para outras máquinas na web. A ideia de combinar conteúdos e link para detectar spam link é considerada razoável, mas a sua abordagem ainda é muito timida. Castillo et al. [Castillo *et al.*, 2007a] também propõem uma ideia semelhante, considerando tanto conteúdo como link para detectar características indiciadores de spam.

'Blog spam' ou 'comment spam' é mais uma técnica maliciosa em que os spammers adicionam links em blogs ou wikis. Em 2005, empresas proprietárias de motores de pesquisa, tais como Google, Yahoo! e MSN,[Wu, 2007] propuseram a utilização de um novo atributo, designado como 'no-follow' (não me siga), para identificar esses intrusos. Esse atributo permitiria que os motores de pesquisa ignorassem esses links. Assim, o efeito do spam em comentários seriam anulados.

Modelos de linguagem, relacionados com a construção sintáctica e morfológicas das frases e palavras podem também ser identificadores de spam. Mishne et al. [Mishne, 2005] acreditam que a linguagem dos 'comment spam' é diferente da linguagem dos blogs ou wikis. Eles propõem identificar spam pela análise de discrepâncias dentro de cada modelo.

Também utilizando técnicas de 'Machine learning' se estuda a possibilidade de detectar comment e blog spam. Kolari et al. [Kolari *et al.*, 2006a;b] propõem a utilização de métodos de aprendizagem automática, como SVM (Support Vector Machines), para detectar 'spam blogs' entre páginas da blogosfera.

'Link bombing' é também uma técnica de link spam. Muitas páginas apontam para uma determinada página, usando o mesmo texto âncora, com o objectivo de inflacionar o seu ranking quando a consulta seja igual a esse texto âncora. Este texto é muitas vezes do tipo 'slang text'[15].Adali et al. analisou este tipo de ligações em [Adali *et al.*, 2005] e provou que o uso deste tipo de palavras no texto âncora pode ser de certa forma um código para responder a queries específicos usados por grupos. Wu [Wu, 2007] refere que mesmo que se identifique este tipo de ataque como 'link bombing', de facto não passa de um 'link farm'.

[15]Abreviaturas nas mensagens de texto, no Twitter, no Facebook, Acrónimos, símbolos do teclado, palavras sem algumas letras são considerados com 'slang'

4.3.4 Técnicas para combater 'page-hiding'

Verificámos já que, para além do content e link spam, uma outra grande categoria existe na classificação do spam: 'page-hiding' ou técnicas de camuflagem e redireccionamento.

Najork apresentou uma patente [Najork, 2003] para detectar páginas camufladas. Ele propôs a instalação de uma barra de ferramentas configurada por forma a permitir que essa barra de ferramentas enviasse a assinatura de todas as páginas visitadas para os motores de busca. Comparando a assinatura enviada (a que foi realmente vista no browser) com a constante nos motores de pesquisa, resultante do 'crawling', pode verificar-se se há ou não coincidência, ou seja, a existência de conteúdo escondido ao crawler. Não está claro se esse método pode distinguir rapidamente se os conteúdos são propositadamente alterados ou se a diferença se deve a páginas geradas dinamicamente e em tempo real.

Cafarella e Cutting [Cafarella & Cutting, 2004] referem as penalizações que os motores de pesquisa podem aplicar a actos de cloaking, quando devidamente confirmados.

Chellapilla and Chickering [Chellapilla & Chickering, 2006] estudaram o grau de cloaking para duas categorias: popularidade e monetarismo. Concluem que cerca de 73% das URLs que usam 'Cloaking' são spam. Da mesma forma concluiram que das páginas que usam cloaking relacionados com incrementos monetários: 98% são spam.

Strider [Yi-Min Wang & King, n.d.] é um sistema para identificar spammers analisando redireccionamentos em páginas da web. Uma pré-lista de páginas identificadas como spam é usada como ponto de partida. Esta lista pode ser expandida adicionando-lhe mais URLs que concorram com qualquer URL desta lista como guestbooks, foruns, web sites, etc. Este processo itera até que a lista convirja. Um dos constrangimentos são os falsos positivos. Para filtrar a lista deste incómodo é lançado um navegador para visitar todos os URLs na lista. O redireccionamento é gravado e os destinos mais visitados após o redireccionamento serão marcados como spam sites. Todos os sites associados a esses agora marcados, também serão marcados como spam.

A maioria dos trabalhos sobre spam assume que qualquer colecção (recolhida pelos crawlers, conforme referimos em 2.1) é pré-processada para remoção de spam antes da indexação. Este resultado não só melhora

a qualidade, mas também reduz o tamanho do índice e de subsequentes recrawls [Attenberg & Suel, 2008].

A avaliação de Josh Attenberg e Torsten Suel [Attenberg & Suel, 2008] efectua uma abordagem orientada para a query de consulta[16], em que os resultados devolvidos por um motor (com a sua própria detecção de spam já aplicadas) são filtradas para remover o restante spam. Ainda que esta acção não diminua o tamanho do índice, tem a vantagem de se concentrar sobre as páginas que realmente aparecem como resultados de buscas típicas.

Este estudo surge como complemento do trabalho de Mishne et al [Mishne, 2005], que usa as estimativas de máxima verosimilhança suavizante para construir modelos probabilísticos exactos da linguagem em uso.

Têm-se verificado nesta área de tarefas de modelação de linguagem, vários trabalhos. É o caso dos estudos de Hearst [Hearst, 1994; 1997], para encontrar sub-topicos dentro de um bloco de texto, estudando multiplas ocorrências de sequências de caracteres. Especificamente, através da comparação de termos comuns nos blocos adjacentes de palavras, poderemos medir a similaridade. Blocos com baixa similaridade são pensados para terem diferentes tópicos.

Spam e Page Rank, como temos vindo a analisar, começam a ser indissociáveis, por isso Du et all. [Du *et al.*, 2007] estudaram a forma como uma vertente de 'Spam Farm' pode ser usado para aumentar o PR, no sentido de determinar as forças e as fraquezas deste classificador.

As novas oportunidades são sempre aproveitadas pelos spammers. Mais um exemplo advém das facilidades de 'tagging' ou seja de o utilizador poder colocar informação, usando ferramentas que crescem de popularidade. Em 'Combating spam in tagging systems' [Koutrika *et al.*, 2007] analisam a forma de combater spam nesta vertente em que o spam pode ser introduzido via comentário.

Também os blogs são cada vez mais populares e mais usados como meios de comunicação social, mas também de outras formas de comunicação. A presença de splogs (spam blog) degrada resultados da pesquisa, bem como desperdiça recursos de rede, criando mesmo a dúvida sobre até que ponto devem ser considerados como páginas, no seu conceito clás-

[16] query-oriented spam detection

sico. [Lin *et al.*, 2007] exploram uma dinâmica temporal, comparativa, para detectar splogs.

Recuperando a ideia de maior vulnerabilidade dos algoritmos de PageRank e Hits aos ataques aos links, Wu et all. [Wu & Chellapilla, 2007], estudam a possibilidade de, seguindo o conceito de hiperligação identificarem comunidades de spam, partindo de um elemento identificado como tal. Esse estudo inicia-se com a plantação de uma semente e simula um 'passeio aleatório' no grafo da web. Esse passeio pretende, de forma tendenciosa, analisar a vizinhança através do uso de probabilidades de deterioração. Truncados os nós mais visitados é produzida uma lista, em ordem invertida que representará uma comunidade de spam.

Saito et all [Saito *et al.*, 2007] estudam a estrutura e distribuição de link farms num grafo de grande escala do JapaneseWeb com 5,8 milhões de sites e 283 milhões de ligações. Encontraram cerca de 0,6 milhões de sites de spam em torno do núcleo. Foi ainda identificado um elevadíssimo número de sites contendo spams de alguma forma ligados a esses núcleos.

Analisámos muitos outros estudos científicos que pretendem, de alguma forma, participar no processo de melhoria de detecção de spam, identificando-o e tornando público os procedimentos para a sua identificação. Muitos trabalhos abordam questões relacionadas com outras formas de spam, nomeadamente as que usam o correio electrónico como forma de disseminação, mas que, neste trabalho, não aprofundámos.

Capítulo 5

Experiências

5.1 Introdução

Neste capítulo expõe-se a forma como se desenvolveram a investigação e as experiências realizadas, começando por efectuar uma pequena referência história aos projectos de investigação sobre WEBSPAM, realizados na Europa, desde 2006.

WEBSPAM - LIP6 - 2006

A primeira recolha de páginas para investigação no âmbito do projecto que analisa as características do WebSpam foi desenvolvida na Universidade de Paris (lip6) [WEBSPAM-UK2007, 2007e], conhecido como 'Web Spam Challenge'.

O principal objectivo do 'Web Spam Challenge' é o de identificar e comparar métodos de 'Machine Learning' (ML) nos procedimentos automáticos de classificação de dados em grandes volumes de informação, agrupados em grafos. Mais precisamente, este projecto, pretendeu extrapolar a classificação para todo o universo do grafo, partindo de apenas uma parte analisada e previamente classificada.

Este grafo compõe-se de um conjunto inter-ligado de 5000 páginas Web. Cada página da Web é rotulado como 'Spam', 'Não é spam' ou 'Borderline[1]' - a última categoria corresponde a uma página da Web em que o conteúdo é apenas parcialmente spam.

[1]Fronteira

WEBSPAM - UK - 2006

A segunda colecção, referenciada como WEBSPAM-UK-2006, foi compilada pela Universidade de Roma 'La Sapienza' e a Universidade de Milão, e encontra-se hospedade no Yahoo! Research Barcelona [Castillo, 2008], que a disponibiliza publicamente.

Este *corpus* consiste em 77 milhões de páginas distribuídas por 12,000 servidores, classificadas ao nível do servidor.

Cerca de 3,000 servidores foram classificados manualmente por, pelo menos, dois colaboradores humanos, com as categorias de 'Spam', 'Not Spam' ou 'Borderline'. Acrescem a estes mais 3,000 servidores que foram automaticamente rotulados como 'Not Spam' em virtude de pertencerem a domínios de confiança, dos quais se salientam: .gov.uk e .police.uk .

Na figura 5.1, podemos observar uma vista parcial do grafo referente ao corpus UK2006:

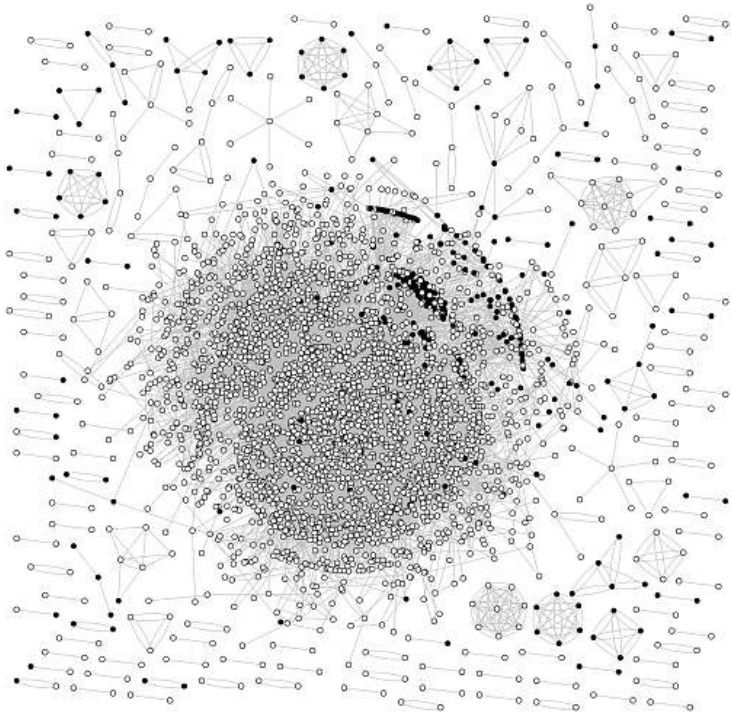

Figura: 5.1: WebSpam UK 2006

A máxima garantia de classificação correcta dos servidores é fundamental para os pesquisadores, e, consequentemente promove a confiança nos resultados das pesquisas por eles efectuada, usando algoritmos de IR.

Infelizmente, ao contrário destas colecções que se encontram controladas e rotuladas, a grande maioria da imensidão de páginas e servidores da Web não se encontram certificados por ninguém[Castillo *et al.*, 2007b].

Se por um lado esta liberalização da Internet foi um dos principais motores do seu sucesso rápido, também é a principal fonte de problemas e desafios para os métodos de IR, principalmente nas tarefas de recolha, indexação, filtragem e 'ranking' nos subconjuntos do Universo onde se verificam manipulações maliciosas, que, como temos vindo a referir, são conhecidas como **spamdexing**.

5.2 Projecto Web Spam 2007

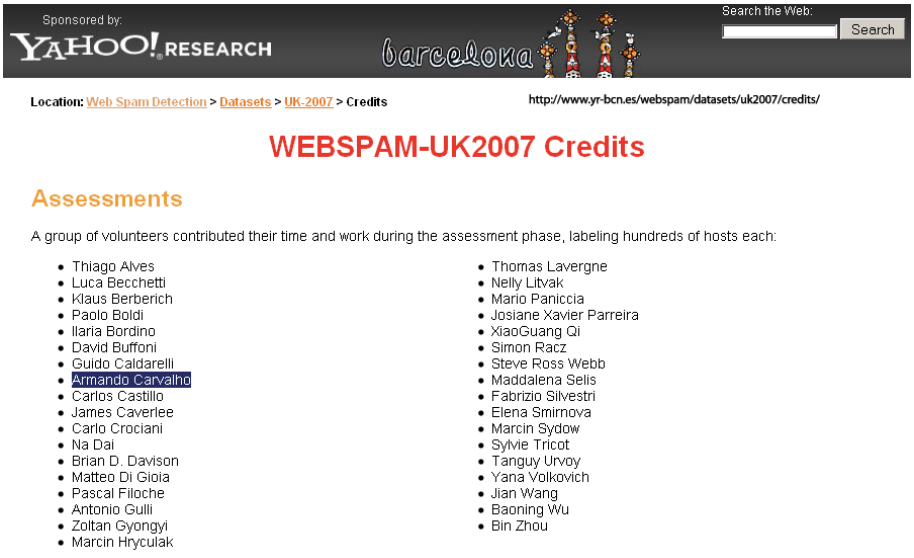

Figura: 5.2: AirWeb 2008 - Creditos

Também com a minha participação no projecto Web Spam UK2007(Spam labeling), Figura 5.2 e Figura 5.4 realizou-se um trabalho de campo, com a mesma tarefa referida anteriormente, a de anotar um largo número de

servidores, com exactos rótulos que indiquem a presença de spam nessas máquinas. Os rótulos definidos para essa classificação foram: 'Spam', 'Borderline', 'Don´t Know', ou 'Normal'[Castillo *et al.*, 2008a] (Figura 5.39).

Figura: 5.3: AirWeb 2008 - Apresentação de Carlos Castilho

Concretamente pretende-se analisar se os algoritmos de classificação de spam reagem de forma positiva em relação à classificação humana. Para isso foram distribuídas, de foma aleatória, um vasto número de páginas, para que, cada host, fosse analisado e classificado por dois investigadores. As suas avaliações individuais foram posteriormente comparadas entre si e, em caso de dúvida, foi possibilitada a revisão da primeira classificação.

Para a equipa do Web Spam UK2007 a pergunta que deve ser colocada durante a classificação é: '*are there aspects of this page that are mostly to attract and/or redirect trafffic*' [pag do WebSpam 2007], no sentido de ser validada a definição para Web Spam de Gyöngyi e Mo-

Figura: 5.4: AirWeb 2008 - Apresentação de Carlos Castilho (Referência aos colaboradores)

lina (2005) '*any deliberated action that is meant to trigger an unjustifiably favorable [ranking], considering the page's true value*' [Gyongyi & Garcia-Molina, 2005]

Para tanto foram definidas em [WEBSPAM-UK2007, 2007d] as linhas de orientação (Guidelines) a utilizar no momento da classificação realizada dentro da plataforma [WEBSPAM-UK2007, 2007a] desenvolvida pela equipa do projecto.

5.2.1 Definição do estudo

A meta indicada foi a de classificar pelo menos 400 hosts por cada um dos investigadores, no sentido de ser possível atingir o objectivo de 400 x 50 = 8,000 rótulos. Considerando que cada host será classificado por dois investigadores conseguir-se-á anotar 4,000 hosts do universo de 100,000 existentes na plataforma de estudo.

A fase de avaliação teve lugar entre 22 de Novembro e 13 de Dezembro de 2007.

Numa segunda fase entre o dia 14 e o dia 21 de Dezembro, foi mostrado, para cada um dos host rotulados por cada investigador, a classificação atribuída pelo outro investigador que também classificou esse host.

Nessa segunda fase foi possível rever as classificações efectuadas durante a fase de avaliação.

5.2.2 Normas de classificação definidas para o WEBSPAM-UK2007

Com o objectivo de uniformizar o mais possível a actuação de cada um dos assessores, necessidade fundamental como iremos analisar em detalhe mais à frente (Secção 5.3), a equipa responsável pelo projecto, definiu um conjunto de regras de classificação dos servidores, ilustradas com casos reais. Como principais ideias a que deveríamos estar atentos ressaltam as seguintes:

- Inclui aspectos concebidos para atrair ou redireccionar tráfego;
- Quase sempre têm intenção comercial;
- Raramente oferecem conteúdos relevantes para os utilizadores.

Os aspectos mais indiciadores de Web Spam são:

- Incluem muitas palavras-chave e links sem relação;
- Usam muitas palavras chaves e muitos sinais de pontuação no URL (Exº.: ../../..);
- Redireccionar o utilizador para uma página que não tem relação, quanto ao conteúdo, com a página inicial;
- Criar muitas cópias da página original com conteúdo duplicado, ainda que, por vezes, apresentando algo de novo em relação à página original.

5.2.2.1 Exemplos práticos de classificação

Conforme definido em 'Guidelines for WEBSPAM-UK2007' [WEBSPAM-UK2007, 2007d], para os efeitos de classificação, poderemos considerar os seguintes casos típicos:

5.2.2.2 Casos típicos de Web Spam

- Páginas cheias de palavras avulsas, ainda que possam representar algum conteúdo, às quais no final, normalmente, se juntam links (Fig 5.5).

- Página cheia de links e palavras avulso sem qualquer significado (Fig 5.6).

- O arquivo de uma mailing-list, copiado apenas para produzir mais palavras (Fig 5.7).

- Técnicas de esconder o texto, como sejam as técnicas de WOW. Fundo da Página: Texto escondido (Fig. 5.8) no final da página que passa despercebido aos humanos.

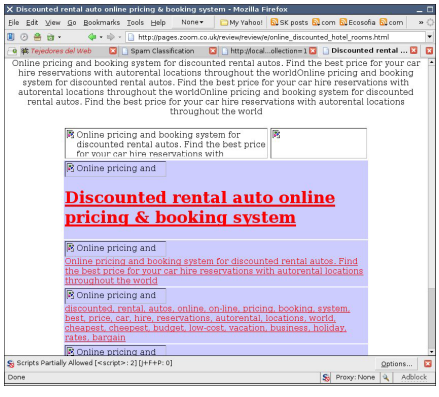

Figura: 5.5: Palavras avulsas sem contexto

Figura: 5.6: Palavras e links avulsos

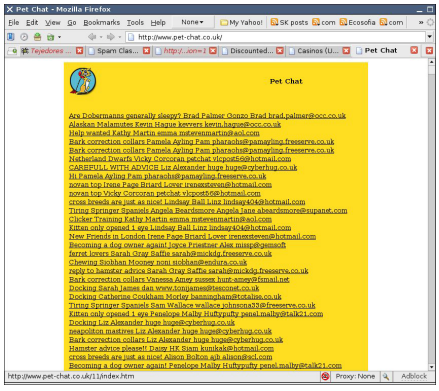

Figura: 5.7: Arquivo de mailing-list

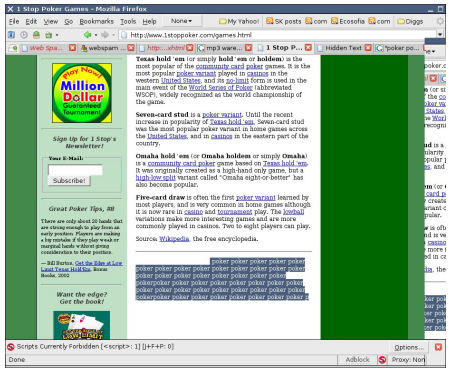

Figura: 5.8: Técnicas de WOW

- Páginas que contém conteúdo gerado automaticamente, mesmo que incluindo algum conteúdo útil (Fig. 5.9)
- Páginas geradas automaticamente com o sentido de servirem de *Parked Domain*,ou seja, um URL preparado para receber eventuais erros de digitação dos utilizadores, que apenas fornecem publicidade. No exemplo pretendia-se o link 'Crat and barrel', e normalmente são páginas que apenas contém links e listas de publicidade (Fig. 5.10)
- Páginas geradas automaticamente com diversos erros de escrita de palavras chave (Fig. 5.11)

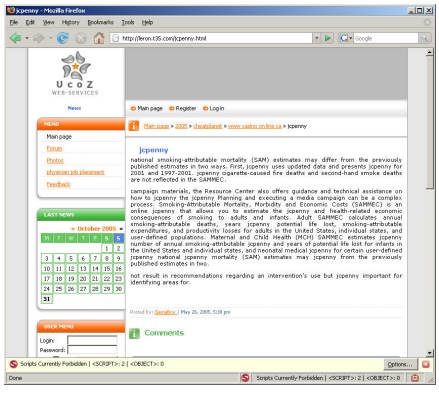

Figura: 5.9: Gerada automaticamente

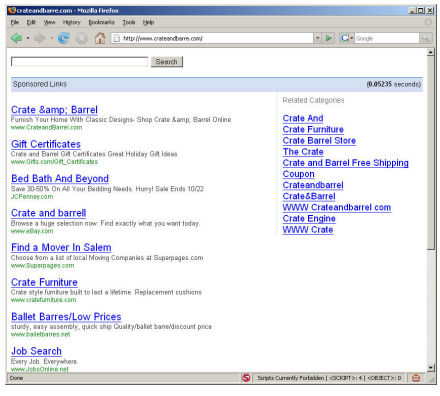

Figura: 5.10: Apenas um Park Domain

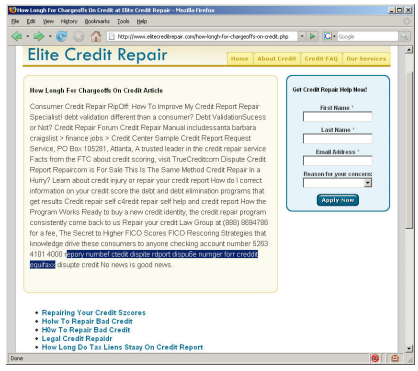

Figura: 5.11: Palavras com erros

- Páginas que apenas contém publicidade, com muito pouco conteúdo. Incluem-se as páginas criadas automaticamente com o fim de vender publicidade (Fig. 5.12)
- Publicidade com indicadores de ratio provavelmente auto-gerados; À Direita: Palavras Chave sem significado apenas para serem pesquisáveis pelos motores de busca (Fig. 5.13)
- Publicidade e Links para outras páginas repletas de publicidade (Fig. 5.14)
- Falsos motores de pesquisa que mostram sempre o mesmo resultado, independentemente do que se escreva no query (Fig. 5.15)

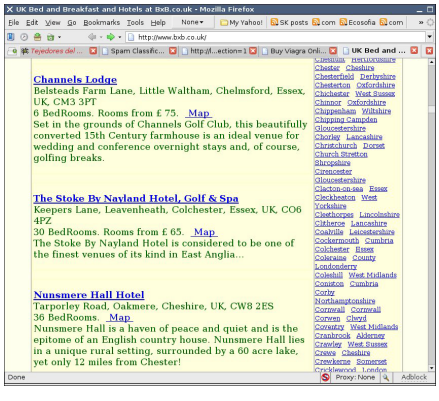

Figura: 5.12: Publicidade sem qualquer conteudo

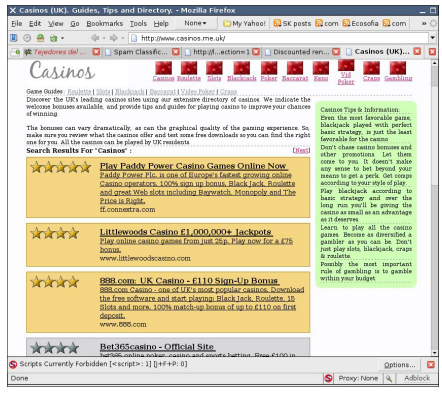

Figura: 5.13: Publicidade e Palavras Chave

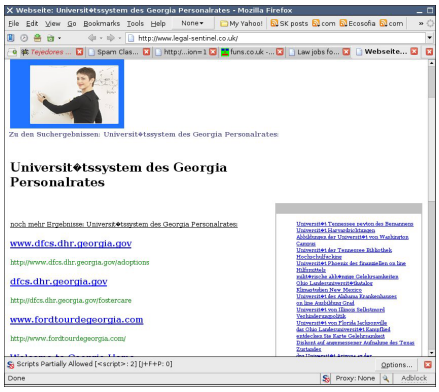

Figura: 5.14: Apenas publicidade e Links

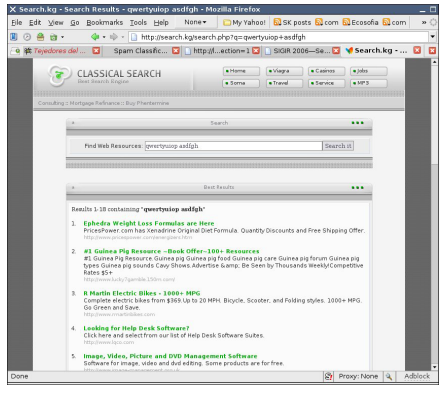

Figura: 5.15: Falso motor de pesquisa

- Classificar também como SPAM os sites que incluam catálogos de produtos que redireccionam para outros sites sem produzirem qualquer mais valia, por exemplo com noticias desactualizadas que foram copiadas de uma qualquer fonte noticiosa (Fig. 5.16).

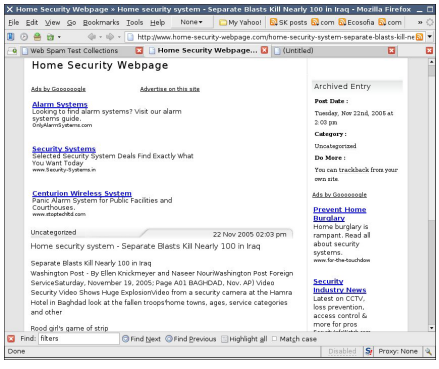

Figura: 5.16: Notícias desactualizadas

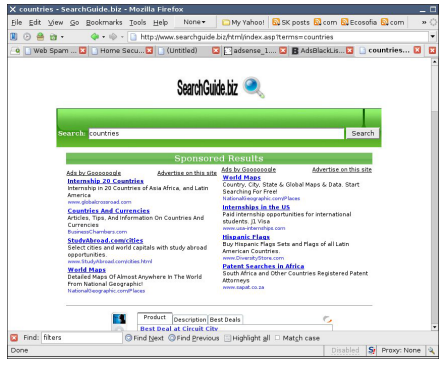

Figura: 5.17: Falso motor de pesquisa

- Falsos motores de pesquisa que apenas mostram publicidade (Fig. 5.17 e 5.18)

- Links gerados automaticamente; Com linhas de texto (provavelmente copiada) a respeito do tópico, e uma página repleta de publicidade (Fig. 5.19)

- Página cujo conteúdo é totalmente publicidade (Fig. 5.20)

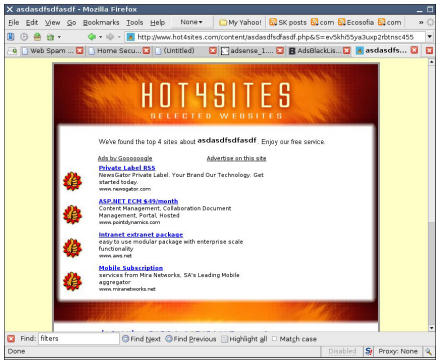

Figura: 5.18: Motor que só mostra publicidade

Figura: 5.19: uma página repleta de publicidade

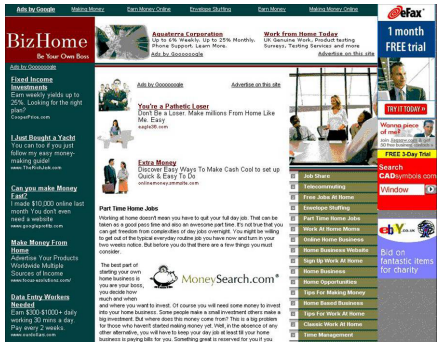

Figura: 5.20: De uma forma geral todo o conteúdo é publicidade

- Paginas que redireccionam automaticamente para outra página, não relacionável, i.e uma página completamente diferente do que é expectável baseado na URL, no texto âncora e / ou nos resultados da pesquisa (Fig. 5.21).

- Páginas com possibilidade de *scripting* (Fig. 5.22)

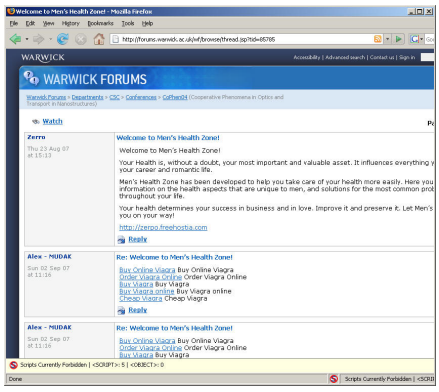

Figura: 5.21: Topo e Direita: Publicidade

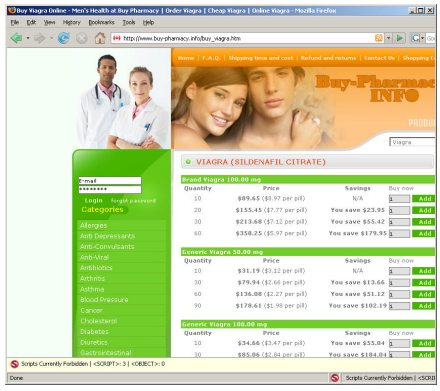

Figura: 5.22: Página com possibilidade *scripting*

- Páginas com links não relacionáveis ou que efectuam trocas de links com demasiados outros parceiros não comparáveis. Poderá haver casos em que, por apresentarem algum conteúdo, podem ser classificados como Borderline (Fig. 5.23).

- Links para sites desenvolvidos pela própria empresa. Também poderemos considerar como borderline, dependendo do ratio de conteúdos originais vs links repetidos (Fig. 5.24).

- Ligações para sites sem qualquer relação (Fig. 5.25).

- Ligações repetidas para páginas do proprio site (Fig. 5.26).

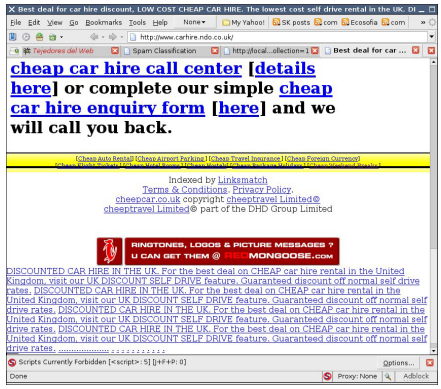

Figura: 5.23: Topo: Palavras e links; Fundo: só links

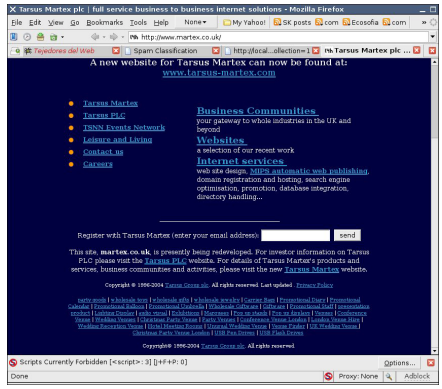

Figura: 5.24: Links para sites desenvolvidos pela mesma empresa

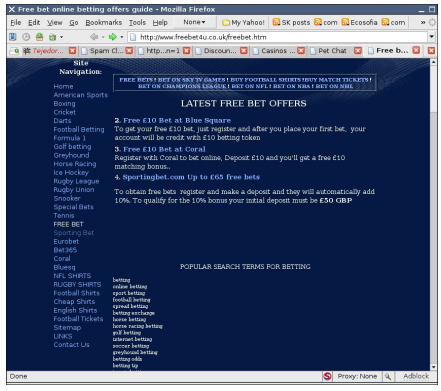

Figura: 5.25: Ligações para sites sem qualquer relação

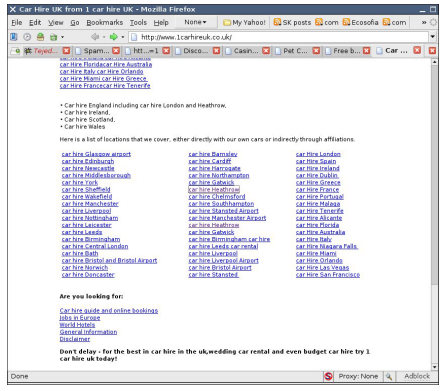

Figura: 5.26: Ligações para páginas do proprio site

- Páginas com conteúdos não relacionados / spam, i.e os seus in-links são provenientes de outras páginas que disponíveis áreas com possibilidade de introdução de conteúdos de forma indiscriminada (pouco fiável), como sejam por exemplo os blogs foruns, guestbooks, etc. Já referimos este exemplo como Spam-oriented blogging em 4.2.1.8

Figura: 5.27: Um exemplo de spam incluído num guestbook. As páginas aqui apontadas são, seguramente, spam.

5.2.2.3 Borderline - Servidores que se encontram na fronteira entre SPAM e NORMAL

- Páginas altamente optimizadas para motores de pesquisa ou preparadas para vender publicidade, ainda que algumas possam oferecer algum conteúdo (Fig. 5.28).

- Site pornográficos optimizado. É, no entanto, aconselhado a que nem todos os sites com pornografia sejam considerados spam. Para este efeitos só são spam se utilizarem truques de spam (Fig. 5.29).

- Algum texto, tudo o resto é publicidade. O texto pode ser roubado da Wikipedia. Se for possível confirmar que o texto é roubado de outro site, considera-se spam. Se não for possível a confirmação deve considerar-se Borderline (Fig. 5.30).

- Parágrafos de texto mas tudo o resto é publicidade. Ainda que o parágrafo possa conter algum valor, é possível a classificação como spam; no entanto se houver duvida do ratio conteudo vs publicidade pode ser classificado como Borderline (Fig. 5.31).

- Páginas a oferecer serviços de optimização por, com troca de links de forma aleatório, participação em grupos de ligações tipo pai-filho, etc. (Fig. 5.31)

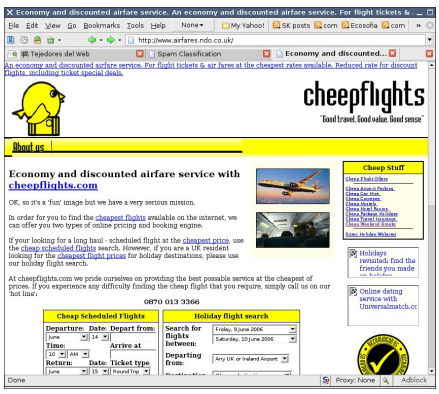

Figura: 5.28: Sites do 'seu grupo de sociedades'

Figura: 5.29: Site pornográfico altamente optimizado

Figura: 5.30: Texto roubado da Wikipedia

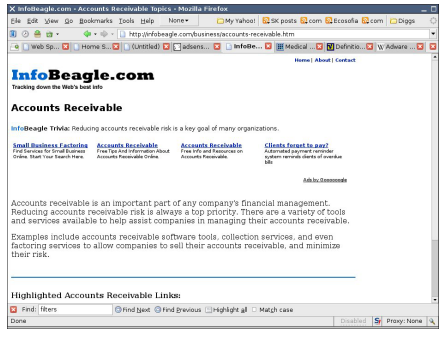

Figura: 5.31: Primeiro parágrafo de texto depois publicidade

- Oferecer serviços de optimização por, com troca de links de forma aleatório, participação em grupos de ligações tipo pai-filho, oferecendo um link para participar num programa de intercâmbio. É muito provável que estas páginas façam parte de link-farms (Fig. 5.32)

- Páginas que oferecem conteúdos original, com alguns links para sites independentes (dependendo da densidade dos links in-bounds pode ser um link-farm (Fig. 5.33).

- Links para vários sites independentes e links para participar em programas afiliados. Isso pode ser normal, borderline ou spam,

dependendo do ratio conteúdo real vs. link (Fig. 5.34)

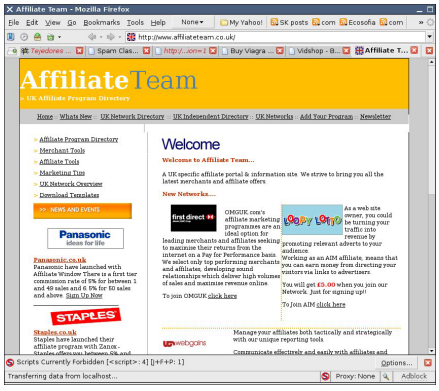

Figura: 5.32: Affiliated

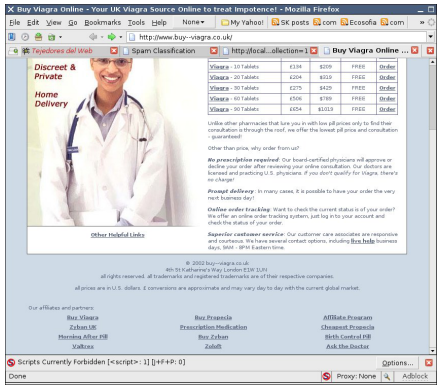

Figura: 5.33: Link Farm e Affiliated

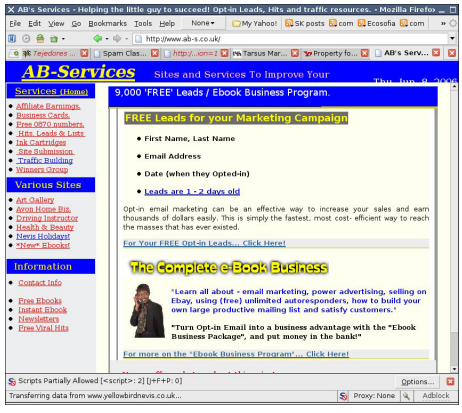

Figura: 5.34: Links para sites independentes e links para participar em programas afiliados

5.2.2.4 NORMAL - Servidores que não contém SPAM

Esta classificação - NORMAL - indicadora de um servidor limpo de spam, refere-se a todos aqueles que não utilizam qualquer um dos truques que temos vindo a indicar como maliciosos - SPAM.

Refira-se que a qualidade gráfica, ou outros pormenores sobre profundidade ou vulgaridade dos conteúdos apresentados não são objecto deste tipo de classificação, pelo que é recomendável que esses factores não influenciem a classificação em causa.

- Sites com conteúdo. Mesmo com títulos grandes, que tem muitas palavras-chave, desde que estejam todas relacionados com a notícia. (Fig. 5.35)

- Forum On-line. Mesmo quando este tenha muitas páginas e pouco conteúdo por página, valida-o o facto de que presta um serviço aos utilizadores (Fig. 5.36)

- Catálogo de shopping válido dado que os produtos são vendidos pelo mesmo comerciante (como o Amazon), ou por vários comerciantes (como o eBay). No entanto, se a página está sempre actuando como uma fachada para redireccionamento para uma loja específica, então é borderline ou spam (Fig. 5.37).

- Directórios on-line. Embora tenham normalmente muitos links não podemos considerar spam se todas as ligações forem cuidadosamente e manualmente seleccionados por editores humanos (Fig. 5.38).

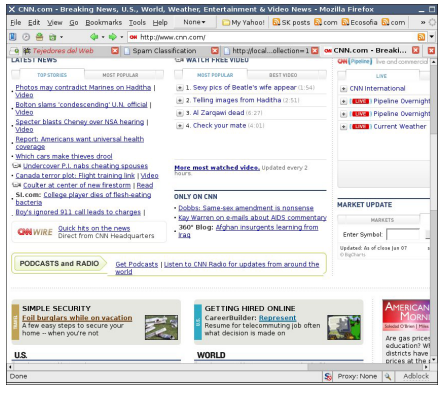

Figura: 5.35: Site com conteúdo

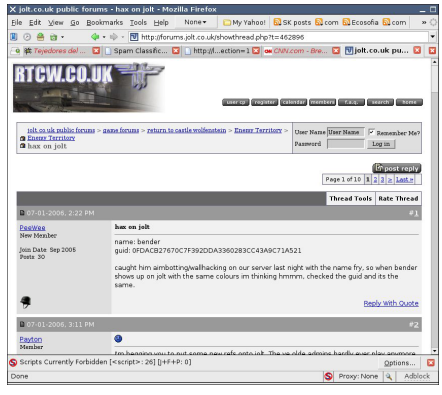

Figura: 5.36: Forum On-line

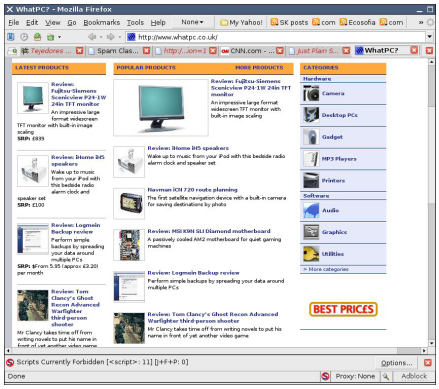

Figura: 5.37: Catálogo de shopping

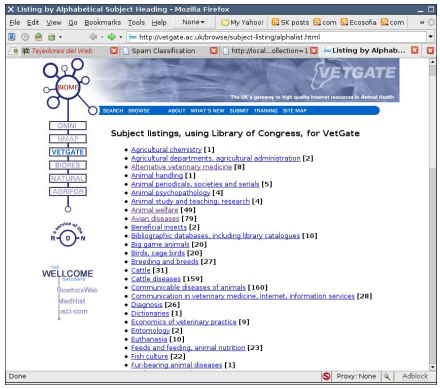

Figura: 5.38: Directório on-line

5.2.2.5 NÃO CLASSIFICADA

Serão identificadas com '???' a que corresponde a etiqueta de NÃO CLASSIFICADA todos os servidores em que não se consiga aceder às suas páginas ou se obtenha o erro 404 de html.

Prevenindo a hipótese de não ser possível enquadrar a classificação em nenhuma das três hipóteses anteriores é apresentada uma quarta hipótese: 'DON´T KNOW'[2]. Será considerada como uma avaliação NULA e não será considerada nos cálculos de rotulagem do servidor.

Esse recurso, de uso muitissimo reservado, pode ser usado para situações como as seguintes:

- Sítios que não consegue classificar.
- Sítios para os quais o conteúdo em cache (interface) e as páginas aleatória mostram apenas páginas em branco, ou apenas '40X não encontrado'.
- Sítios a que não seja possível aceder ou que necessitem de password.

[2]Na interface aparece com o indicador ???

5.3 A importância da unanimidade na classificação humana de servidores

Para complementar o trabalho referido na seccção anterior (5.2), elaborámos um estudo sobre o comportamento dos assessores em dois momentos distintos: (**1**) Na fase que foi designada por INITIAL e que compreendeu não só a primeira classificação, mas eventuais alterações efectuadas dentro do prazo estabelecido, e (**2**) na fase designada por REVISION em que são apresentados a cada um dos intervenientes as classificações que ele próprio efectuou e todas as outras classificações, de outros intervenientes, para esses mesmos sites.

Mantém-se como um dado adquirido que, quem classifica e tem que decidir, adopta como principal critério de decisão, nos casos fronteira entre spam e não spam, a percepção do esforço despendido pelos autores das páginas web para proporcionar bons conteúdos, contra o esforço despendido na tentativa de pontuação elevada em motores de busca.

O braço de ferro aqui produzido, assim denominado por Castilho [Castilho *et al.*, 2006], criou, por sua vez, **o domínio do Contraditório nos procedimentos de IR**[3] que estuda a forma de adaptar as técnicas para recuperação de informações para contextos em que, parte da colheita, tenha sido modificada maliciosamente com o único objectivo de afectar o ranking dos algoritmos.

Esta decisão de tornar clara a definição sobre se estamos ou não perante spam, é uma tarefa difícil do ponto de vista matemático e algorítmico, tanto mais que não é clara a manutenção da mesma posição do ser humano quando, depois de rever possíveis classificações, lhe são submetidas para comparação outras posturas, de outros classificadores, também humanos, mas com outra perspectiva sobre o limite teórico do spam.

5.3.1 Trabalho relacionado

Este trabalho tem vários outros trabalhos âncora, onde se reflectem preocupações sobre '*web spam classification*', e principalmente os inte-

[3] Adversarial Information Retrieval

grados nas apresentações efectuadas para os congressos, designados por AirWeb, relacionados com melhorias nos procedimentos de recuperação de informação.

Neste nosso caso particular, apertamos o universo e a incidência específica deste trabalho centra-se sobre o estudo das particulares dificuldades provocadas pela introdução de SPAM nos resultados da selecção levada a cabo pelos motores de busca [Castillo *et al.*, 2007a] e das pesquisas desenvolvidas para depurar e melhorar os algoritmos de detecção de spam nos sistemas de recuperação da informação.

Apenas com a análise e estudo do comportamento humano em situação real, é possível melhorar os algoritmos de computação, principalmente do ponto de vista do *ranking* de classificação final, evitando o mais possível os truques introduzidos como indutores de melhorias de classificação.

Para que seja possível uma resposta com qualidade, que permita acesso o mais directo possível às páginas que realmente interessam, desde o início do WWW, tornou-se necessário classificar as páginas de acordo com a possibilidade de respostas conhecidas à pergunta efectuada aos motores de pesquisa e por estes aos indexadores.

Este trabalho, conforme caracterizado em 'Social Media Research Blog'[4] nem sempre é fácil de obter, principalmente *sem se estudar o comportamento entre os anotadores*:

> **The Cost of Manual Annotation**
>
> For many machine learning tasks it can be quite difficult to get the "ground truth" data. In some cases the best way to verify the results is by a painful, laborious and mindnumbing task of labeling data manually. When Pranam and I were working on the Splog detection task, we spent a good deal of time painstakingly labeling independently if a blog from a random sample was legitimate or spam. The part that makes this task difficult is:
>
> - Sometimes it is not clear when something in a blog is a splog

[4] http://socialmedia.typepad.com/blog/2008/06/the-cost-of-manual-annotation.html

- You need to also look at the inlinks and outlinks
- Plagiarized content makes it harder to judge authenticity
- Sploggers are getting more sophisticated in the methods they are using

On an average we spent about 2-3 minutes per blog and in the end, were only able to hand label a small collection. In comparison to many other tasks this was still a relatively straightforward judgment. Consider the task of relevance ranking that NIST has to perform each year for the TREC tracks. Here the goal is for the annotators to figure out if a result is relevant for a query. The guidelines are strict and NIST has many professional annotators who are trained to perform these tasks. Even more complicated are some of the annotation that might be required in certain Natural Language Processing experiments. These can range anywhere from just verifying a parse tree or an output to actually constructing gold standards or hand crafted parse trees. Moreover, some NLP tools require tremendous amounts of linguistic resources – be it tediously constructing an ontology, lexicon, gazetteer lists or identifying word senses. Many of these tasks require linguists or experts whose time might be quite valuable.

Lets consider a simple case where the annotator was asked to label a URL with a tag. Lets also say that it takes roughly a minute to load the page, quickly glance over it, make a judgment and then type in the appropriate labels. I know from experience that this is not a minute but more like 1.5-2 minutes on an average (try it! it is a braindead boring task and if you are asked to do it continuously, you will slow down!). If say I can work 10 hours on this task without loss in quality of my annotations, it would only result in 600 URLs being tagged. UMBC pays lets say around $10/hour for on-campus jobs. That means we would spend about $100 just to label 600 URLs. Not so sure if that would be the best way I would like to spend a hundred bucks! Additionally, just one human annotator is never sufficient. You always need to answer questions like : **'So, what was your inter-annotator agreement?'**. Well then you just blew another $200 or $300 on this task and still have just 600

> URLs marked up. No wonder del.icio.us and Flickr are such amazing sources for free (yaay!), human assessed labels and annotations. It works out great if you can use these instead.

5.3.2 Web Spam Detection

'*The significant opportunities for new business models come at the cost of a high reward, and low bar (in terms of cost), for unscrupulous players hoping to tilt the game in their favor, at the expense of the common good*'[Svore *et al.*, 2007b].

As técnicas de detecção de spam são normalmente concebidos para determinados tipos conhecidos de Web spam e são incapazes, ou ineficientes, para as recém surgidas variantes de spam, como refere Liu [Liu *et al.*, 2008]. Diferentes algoritmos matemáticos tentam aprender - utilizando técnicas de análises preemptivas como o algoritmo Bayesiano [Bíró *et al.*, 2008; Liu *et al.*, 2008], ou o recente algoritmo designado por WITCH [Abernethy *et al.*, 2008a] - como detectar irregularidades classificáveis como spam.

O exemplo dado por Benevuto et al. [Benevenuto *et al.*, 2008] no seu estudo sobre a análise dos problemas na detecção de 'video spammers' é um indicador interessante da evolução que está a acontecer nesta área.

Com a participação no projecto 'WEBSPAM-UK2007 (Spam labeling)', teve lugar um 'trabalho de campo' em que a tarefa específica foi a de classificar um largo número de servidores, **com etiquetas precisas sobre a existência de spam nessas máquinas**. As etiquetas definidas para esta classificação foram: 'Normal', 'Spam', 'Borderline' ou 'Couldn´t classify'(Figura 5.39)

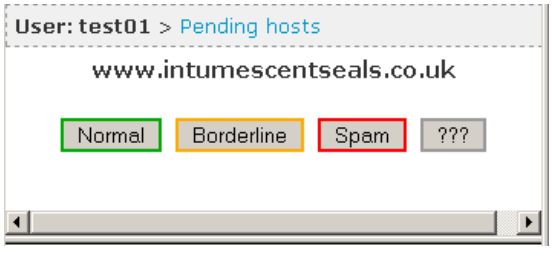

Figura: 5.39: Classificadores dos servidores possíveis de ser utilizados nas fases inicial e de revisão

Especificamente pretendeu-se examinar, por comparação, se o comportamento dos algoritmos de spam reagem em concordância com a classificação efectuada por humanos [WEBSPAM-UK2007, 2007e]. Com esse fim foi distribuído um largo número de páginas com o cuidado de garantir que cada servidor fosse classificado por, pelo menos, dois desses assessores humanos. No final do período definido para o projecto, foi possível reanalisar, por cada um dos assessores, as classificações em que não houve concordância com outros elementos.

Para a equipa do WEBSPAM-UK2007, a questão que pretendiam fosse sempre equacionada, no exacto momento da classificação, era: 'existe algum aspecto nesta página que possa **atrair ou redireccionar tráfego**' [WEBSPAM-UK2007, 2007d], no sentido de irem de encontro à definição de Gyöngyi and Molina [Gyongyi & Garcia-Molina, 2005] sobre web spam: 'any deliberated action that is meant to trigger an unjustifiably favorable [ranking], considering the page´s true value'

As orientações para o WEBSPAM-UK2007 estiveram definidas no site oficial do evento [WEBSPAM-UK2007, 2007d], e os assessores, para análise e classificação, utilizaram uma plataforma concebida para o efeito [WEBSPAM-UK2007, 2007a].

A fase de avaliação e classificação decorreu entre 22 de Novembro e 13 de Dezembro de 2007, conforme referido em 5.2.1.

Na segunda fase, que decorreu entre 14 e 21 de Dezembro, foi dada a possibilidade a cada assessor de rever a sua classificação, sendo-lhe possível visualizar as classificações de outras pessoas, mas apenas referente aos servidores que cada um classificou.

5.3.3 Descrição das bases de dados

Foi usada a versão pública da base de dados do WEBSPAM-UK2006, como Castilho [Castilho *et al.*, 2006; Castillo *et al.*, 2007a] e a colecção de WEBSPAM-2007, que na opinião de Bíró [Bíró *et al.*, 2008] é muito mais sensível a aspectos de conteúdo do que a características dos links. Esta segunda base de dados também está disponível para acesso público na www [WEBSPAM-UK2007, 2007c].

Ambos foram baseados num conjunto de páginas obtidas a partir de

um rastreamento[5] do domínio .uk. O conjunto de dados foi obtido em Maio de 2006 e Maio de 2007, respectivamente pelo grupo de investigação do 'Laboratory of Web Algorithmics da Università degli Studi di Milano', de acordo com as referências de Castilho [Boldi *et al.*, 2002; Castillo *et al.*, 2007a].

No final dos procedimentos, no caso do WEBSPAM-UK2006, 6,552 servidores foram classificados [Castillo *et al.*, 2007a]; na colecção WEBSPAM-UK2007, que incluia 114,529 servidores foram classificados 6,479 desses servidores [WEBSPAM-UK2007, 2007c].

Analisaremos a primeira parte (SET 1), que inclui 4,275 servidores, uma vez que representa o universo, conforme a confirmação por mail de Carlos Castilho:

> 'Set1 e set2 são apenas dois subconjuntos, aleatórios, dos dados marcados pelos assessores. Se houver qualquer diferença entre os dois deles, terá sido mero acaso'

A separação do universo de servidores classificados em 2 Set's foi de aproximadamente 2/3 para treinos da comunidade científica e 1/3 ficou reservado para testes.

> *For the purpose of the Web Spam Challenge 2008, the labels were released in two sets. SET1, containing roughly 2/3 of the assessed hosts was given for training, while SET2 containing the remaining 1/3, was held for testing.*

Acreditamos, por isso, que o Universo agora analisado é amostra suficiente para o objectivo científico que nos propusemos.

Numa primeira abordagem poderemos comparar os classificadores do WEBSPAM 2006 com os de 2007, onde sobressai que o classificador Normal foi o mais utilizado em ambas, no entanto, em 2007, o valor aumenta mais de 20%, o que pode ser um primeiro indicador da evolução do controlo sobre o spam, mas, por outro lado, pode querer representar que a vulgarização de alguns actos ou a melhoria de alguns truques tornou imperceptíveis páginas com spam (5.19% vs 22.08%). Por aqui abrem-se, desde logo, interessantes perspectivas de análise.

[5] crawler

Por outro lado - e presente o objectivo final de que a classificação deve ser o mais clara possível sobre SIM ou NÃO - o classificador de fronteira (BORDERLINE) desce de 10.82% para 2.39%, o que é excelente do ponto de vista do objectivo referido que implica atingir, no limite, um valor de 0% para as incertezas.

A distribuição das classificações é mostrada na tabela 5.1.

	WEBSPAM-UK2006		WEBSPAM-UK2007 SET 1	
Classificação	**Frequencia**	**Perc.**	**Frequencia**	**Perc.**
Normal	4,046	61.75%	3,776	88.33%
Spam	1,447	22.08%	222	5.19%
Borderline	709	10.82%	102	2.39%
Could not be classified	350	5.34%	175	4.09%

Tabela: 5.1: Classificação dos servidores efectuada pelos assessores.

5.3.4 Distribuição dos classificadores

Do ponto de vista de interesse dos ataques dos spammers um primeiro 'filtro', ou se quisermos do ponto de vista contrário, um primeiro 'alvo' é determinado pelo subdomínio a que o site está ligado.

Interessa por isso saber que, de entre o universo de onde foi extraída a nossa colecção (.uk), o subdomínio mais volumoso é o co.uk, dada a sua implantação e usabilidade pela comunidade comercial do Reino Unido, conforme Tabela 5.2 e Figura 5.40

Subdomain	Quant.	Normal	Spam	Borderline	Unclassified
ac.uk	171	165	–	–	6
bl.uk	2	2	–	–	–
co.uk	3,354	2,916	201	93	144
gov.uk	79	75	-	–	4
ltd.uk	6	6	–	–	–
me.uk	8	7	1	–	–
mod.uk	2	2	-	–	–
net.uk	1	1	–	–	–
nhs.uk	17	15	-	–	2
org.uk	574	527	20	9	18
plc.uk	3	3	–	–	–
sch.uk	58	57	–	–	1
Total	4,275	3,776	222	102	175

Tabela: 5.2: Distribuição dos classificadores por subdomínio.

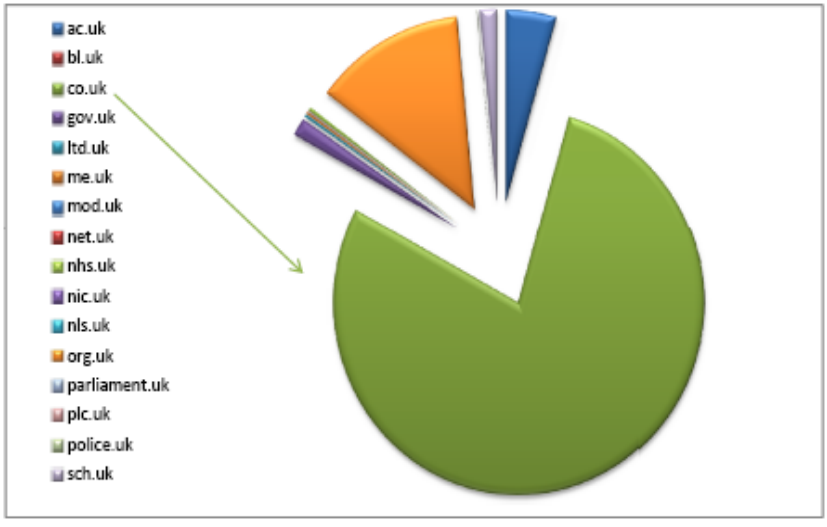

Figura: 5.40: 2007 URL (SET 1) distribution

Os subdomínios mais representados (co.uk e org.uk) apresentam elevados valores para a classificação de Normal, conforme Tabela 5.3 sendo de realçar que o 'co.uk', de características eminentemente comerciais, apresenta valores mais baixos para sites NORMAL (86.94% vs 91.81%) e mais elevados para SPAM (5.99% vs 3.48%).

Mesmo o indicador Borderline é mais elevado para esse subdomínio comercial. É também de referir a paridade com a tabela 5.1, antes referida.

Subdomain	Quant.	Normal	Spam	Borderline	Unclassified
co.uk	3,354	86.94%	5.99%	2.77%	4.29%
org.uk	574	91.81%	3.48%	1.57%	3.14%

Tabela: 5.3: Principais subdominios de distribuição, em percentagem

5.3.5 Representatividade da Amostra Usada

A tabela 5.4 apresenta o universo tratado após a segunda fase de classificação (fase de revisão), versus o universo total de dados (114,524 servidores).

Subdomínio	Dataset	Percentagem	Classif	Percentagem
ac.uk	5,063	4.42%	171	4.00%
bl.uk	18	0.02%	2	0.05%
co.uk	89,953	78.54%	144	78.46%
gov.uk	1,708	1.49%	79	1.85%
ltd.uk	257	0.22%	6	0.14%
me.uk	216	0.19%	8	0.19%
mod.uk	52	0.05%	2	0.05%
net.uk	34	0.03%	1	0.02%
nhs.uk	339	0.30%	17	0.40%
nic.uk	1	0.00%	0	0.00%
nls.uk	1	0.00%	0	0.00%
org.uk	15,141	13.22%	574	13.43%
parliament.uk	6	0.01%	0	0.00%
plc.uk	25	0.02%	3	0.07%
police.uk	57	0.05%	0	0.00%
sch.uk	1,658	1.45%	58	1.36%
Total / Avg	114,528	–	4,275	–

Tabela: 5.4: Representatividade da amostra

O modelo usado na distribuição de servidores pelos participantes, para classificação, assegura as relações entre o universo individual e o

universo do DataSet, pelo que podemos concluir pela validade da amostra.

5.3.6 Experiências e Resultados

Neste estudo analisámos, para melhor compreensão, o comportamento humano relacionado com:

- Qual o grau de concordância em spam / nonspam?
- Quão diferente são pessoas diferentes (algumas usam muito a classificação spam outras evitam essa classificação)?
- Em que grau a opinião das pessoas muda quando lhe são apresentadas as opiniões de outras pessoas, sobre o mesmo assunto?
- A classificação com o rótulo 'borderline' - como se comportam os assessores?

5.3.6.1 Qual o grau de concordância em spam / nonspam?

Na tabela 5.5 analisamos os diversos grupos, ou seja sites completos, após a classificação final, nomeadamente no que diz respeito à forma como os vários assessores que classificaram cada um dos sites se comportaram, quanto à concordância total ou parcial.

	Nível de Concordância		
	Total	Parcial	Classif.
Nonspam	3,504	272	3,776
Spam	157	65	222
Borderline	102	0	102
Unknown	175	0	175
Total	3,938	337	4,275

Tabela: 5.5: Grau de concordância.

A análise desses resultados revela que 92.80% (3,504 de 3,776) dos sites classificados como 'Nonspam' mereceram a concordância total dos intervenientes, enquanto que em apenas 7.20% (272 de 3,776) se verificou não concordância total entre os assessores, o que, à partida, pode ser um indicador do nível dos critérios dos assessores bastante equilibrado.

A utilização do classificador 'Spam' foi muito mais baixo. 70.72% dos humanos concordaram inteiramente, havendo um acordo parcial de 29.28%.

Pensamos que esta análise ficaria muito enriquecida se fosse possível incluir outros elementos para análise, como sejam o país de origem, o grau académico, a idade e o sexo do assessor, de um modo particular nos casos em que não se verifica acordo total.

5.3.6.2 Quão diferente são pessoas diferentes (algumas usam muito a classificação spam outras evitam essa classificação)?

Como referimos no final da secção anterior, é nossa convicção que há muitos factores de personalidade que constrangem a objectividade pedida no momento da atribuição de uma classificação.

Como podemos analisar na Tabela 5.6, os primeiros 10 assessores, por ordem do número total de servidores classificados, de um modo predominante, classificaram como NONSPAM. Excepção pontual para o primeiro em que o uso de BORDERLINE ou UNKNOWN é francamente elevado, o que pode levar a concluir que: (**1**) estes consultores efectuaram uma triagem inicial para validar, desde logo, os servidores em que acreditaram ser confiáveis ou (**2**) que os primeiros possam ser membros da equipa de concepção e que conhecem bem o modelo.

A análise dos dados demonstra que, conforme referido na tabela 5.7, para a classificação **initial**, se verificou uma classificação massiva como NONSPAM (76.30%), existindo também uma percentagem considerável de indecisos (BORDERLINE 6.21%). Estes últimos aproximam-se bastante dos sites classificados como SPAM (6.54%).

Assessor	Spam	Nonspam	Borderline	Unknown	Total
1	14	292	8	8	322
2	24	257	23	16	320
3	24	239	15	21	299
4	6	215	12	50	283
5	7	249	14	11	281
6	5	243	14	18	280
7	16	236	22	6	280
8	22	226	15	15	278
9	12	235	12	15	274
10	6	221	11	34	272

Tabela: 5.6: Contagem das Classificações efectuadas pelos primeiros 10 assessores com o maior numero de sites classificados

Phase / Classif.	Borderline	Nonspam	Spam	Unkn.	Total
Initial	619	7,606	652	1,092	9,969
Revision	186	301	142	35	664
(A) %(Revis./Initial)	30.05%	3.96%	21.78%	3.21%	6.66%
(B) %(Initial/Total Initial)	6.21%	**76.30%**	6.54%	10.95%	-
(C) %(Revis./Total Revis.)	28.01%	45.33%	21.39%	5.27%	-

Tabela: 5.7: Evolução da classificação por fases

5.3.6.3 Em que grau a opinião das pessoas muda quando lhe são apresentadas as opiniões de outras pessoas, sobre o mesmo assunto?

Considerando o objectivo: 'Alteração de classificador', usámos apenas os dados referentes aos servidores classificados na segunda fase (REVISION). Daí obtivemos a primeira classificação por assessor e todas as classificações de outros assessores para os mesmos servidores.

Esta análise permite-nos avaliar o grau de influência dos 'vizinhos' que analisaram o mesmo conteúdo.

Daqui que, dos 9,969 sites acedidos na fase inicial, 664 foram alterados (Tabela 5.7), ou seja, apenas 6.66%.

Deste grupo de 664 servidores, verificaram-se alterações em 61.60% (100.00% - 38.40%), motivadas pela visualização das classificações dos outros assessores (Tabela 5.8).

REVISTA (A)		INICIAL (B)		ALTERARAM (B/A)	MANTIVERAM
SPAM	142	**SPAM**	52	36.62%	7.83%
		NONSPAM	35	24.65%	
		BORDERLINE	36	25.35%	
		UNKNOWN	19	13.38%	
NONSPAM	301	SPAM	76	25.25%	
		NONSPAM	120	39.87%	18.70%
		BORDERLINE	93	30.90%	
		UNKNOWN	12	3.99%	
BORDERLINE	186	SPAM	45	24.19%	
		NONSPAM	74	39.78%	
		BORDERLINE	57	30.65%	8.58%
		UNKNOWN	10	5.38%	
UNKNOWN	35	SPAM	3	8.57%	
		NONSPAM	3	8.57%	
		BORDERLINE	3	8.57%	
		UNKNOWN	26	74.29%	3.92%
	664	-	664	-	38.40%

Tabela: 5.8: Fase de Revisão: Mudança de Classificador

Podemos então tirar uma **primeira conclusão** de que, **em caso de não concordância**, existe uma alta probabilidade (61.60%) de ser mudada a opinião inicial, o que nos pode levar a concluir que o **espírito de isenção absoluto que as medidas de classificação devem ter, não foram totalmente atingidas**.

De entre os que reviram a classificação inicial(18.7%), concluímos que NONSPAM é a classificação menos alterada, e que há uma convergência no sentido desta classificação. Verifica-se uma tendência para dar mais facilmente como provada esta situação de NONSPAM do que o contrário, ou seja, uma **segunda conclusão**, de que **na falta de clara evidência de spam não se usa este classificador**.

Na mesma tabela 5.8 analisamos ainda as proveniências, ou seja, de que forma cada um dos produtos finais recebem informação das diversas outras proveniências

Daqui é também possível concluir que dos 142 sites agora classificados como SPAM, apenas 52 (7.83%) mantiveram a classificação. Mudaram de classificador para SPAM 90 sites com outras classificações anteriores (35 de NONSPAM, 36 de BORDERLINE e 19 de UNKNOWN). É também relevante de assinalar que nem todas as classificações mantidas

(neste primeiro caso as que mantiveram o classificador de SPAM e que são 52) são todas classificações de unanimidade, dado que classificação final resulta daquela que for mais preponderante, i.e. por exemplo se dois assessores classificam de SPAM e um de UNKNOWN o site fica marcado como SPAM.

De facto idêntica interpretação pode ser aplicada a todas os classificadores finais aqui estudados. Calculados esses valores, conforme tabela 5.9 verificámos possíveis novas razões.

	Inicial	Final	Variações	%	
SPAM	176	142	-34	-19.32%	a)
NONSPAM	232	301	69	29.74%	b)
BORDERLINE	189	186	-3	-1.59%	-
UNKNOWN	67	35	-32	-47.76%	c)

Tabela: 5.9: Evolução da classificação por fases:
a) A diminuição da classificação como spam pode ter a ver com o que é conhecido como 'votar vencido', neste caso em NONSPAM;
b) O aumento poderá ter a ver com as razões anteriores
c) A vizinhança ajuda a tomar decisão

De salientar que é possível extrair uma **terceira conclusão**, resultante da alínea c), de que a **vizinhança ajuda a tomar decisão**.

5.3.6.4 A classificação com o rótulo 'borderline' - como se comportam os assessores?

A classificação do tipo 'borderline' é, sem margem de dúvida, uma classificação alternativa, que se pretende seja de utilização muito reduzida, mesmo residual, por forma a que a classificação seja limitada a SPAM e/ou NONSPAM. Daí que, as conversas ocorridas, principalmente na primeira fase, em salas de chat destinadas a esclarecimentos ao grupo de voluntários, eram no sentido de serem criados alertas elucidativos, nos casos de difícil classificação apenas com SPAM ou NONSPAM, que ajudassem na melhoria das regras heurísticas a incluir nos procedimentos matemáticos que viessem a ser desenvolvidos.

Por isso verificámos uma especial preocupação humana na utilização do classificador 'borderline'. Como se pode ver nas análises referidas nas tabelas 5.10 e 5.11, a convergência é absoluta para as classificações de

SPAM ou NONSPAM. Apenas em 7.57% dos casos foi utilizado 'Borderline'

	INICIAL	REVISTA	TOTAL
Spam	652	142	794
Nonspam	7,606	301	7,907
Borderline	619	186	805
Unknown	1,092	35	1,127
TOTAL	9,969	664	10,633

Tabela: 5.10: Evolução das classificações por tipos

	INICIAL	REVISTA	TOTAL
Spam	6.54%	21.39%	7.47%
Nonspam	76.30%	45.33%	74.36%
Borderline	6.21%	28.01%	7.57%
Unknown	10.95%	5.27%	10.60%

Tabela: 5.11: Médias de Classificação por tipos

Na fase inicial (primeira classificação dos consultores, sem qualquer forma de revisão e envolvendo um maior numero de indivíduos) o nível de classificação como 'Borderline' foi bastante baixo: somente 5.47%. Perante estas análises, é-nos legítimo concluir - como referimos atrás - que a opção de 'Borderline' é de uso bastante raro.

E finalmente poderemos avaliar quantos é que foram influenciados conduzindo à mudança de classificador. Conforme a Tabela 5.12 a maioria (61,59%) alteraram a classificação.

MUDARAM	35+36+19+76+93+12+ ...	409	**61.59%**
NÃO MUDARAM	52+120+57+26	255	**38.41%**

Tabela: 5.12: Análise sintética do grau de mudança

5.3.7 Conclusões possíveis

Ressalta como primeira evidência que é muito dificil, do ponto de vista estritamente humano, encontrar uma definição para o que é SPAM e o que não é.

A subjectividade da classificação reside fundamentalmente na interpretação humana do conceito, o que provoca uma área cinzenta à volta da desejável clara definição de SPAM ou NÃO SPAM. Esta fronteira, cujo limite se pretende seja zero, é seduzida pela classificação dos vizinhos, como indicado pelas alterações feitas durante a revisão

Podemos também concluir que a classificação inicial de NÃO SPAM e SPAM são menos voláteis e mais duradouras. Destes dois a classificação mais permanente é a de NÃO SPAM.

Como referência final, pensamos que também seria importante avaliar o tempo que cada assessor demora a efectuar cada tarefa de classificação, e bem assim a profundidade da decisão, avaliável pelo número de Inbounds e Outbounds analisados imediatamente antes da tomada de decisão.

Respondendo à principal preocupação que demonstrámos no princípio do trabalho, podemos concluir que a classificação como 'Borderline' é reduzida e de utilização muito cuidada pelo ser humano e de que o seu aumento, que poderemos conotar como de 'indecisos', por um lado, e o valor de 61.59% (Tabela 5.12) para os que alteraram a sua classificação inicial, são indicadores de que a **vizinhança, para além de influenciar, pode gerar alguma confusão**.

Capítulo 6

Conclusões e trabalho futuro

O que é o Spam? De que modo dificulta os sistemas de recuperação na Web?

Estas foram as principais preocupações, de variadas perspectivas, que nos acompanharam em todo o nosso trabalho de investigação sobre Web Spam.

Primeiro identificámos conceitos: motores de pesquisa, algoritmos de classificação e a sua evolução temporal. Depois, aproveitando o que de melhor conhecemos da comunidade científica, apresentámos uma sinopse de técnicas anti-spam. Como trabalho próprio apresentámos um estudo que acrescenta um factor sociológico a todas as questões estudadas.

Tudo se deve ao grande crescimento da web.

De facto, desde o seu tímido início - com a função de partilha de dados entre Físicos - a web cresceu, e continua a crescer, sendo hoje também um pólo central de cultura, de educação e, acima de tudo, de vida comercial. Milhões de utilizadores executam diariamente transacções financeiras nas suas páginas da web, que podem ser desde compra de utensílios, de livros, de viagens e hotéis, até à gestão de carteiras de aplicações financeiras. É conhecida de todas nós, sobretudo do ponto de vista das acessibilidades, a forma como a web modificou o paradigma da informação.

Devido à espantosa quantidade de informações disponíveis na web, os utilizadores - todos nós - depressa nos habituámos a consultar / pesquisar conteúdos usando os motores de pesquisa. Para cada consulta,

um motor de pesquisa identifica as respectivas páginas na web e apresenta aos utilizadores os links para essas páginas, geralmente em lotes de 10/20 respostas. Em face das ligações, ordenadas por grau de comparação com a pergunta, cada utilizador pode optar por cada uma das ligações fornecidas.

Este 'click' que determina a primeira opção, é efectuado, de uma forma geral, no primeiro grupo de páginas que é fornecida ao utilizador - **85% das vezes, as pessoas apenas consultam os primeiros 10 resultados da resposta** [Metaxas & DeStefano, 2005; Silverstein *et al.*, 1999] -, daí a alta vantagem (no conceito comercial) de quem fica colocada nesse 'top ranking'.

A tentação de colocar, de uma forma mais ou menos clara, as páginas nesse grupo primeiro de selecção, existe praticamente desde que existe Internet [Ntoulas *et al.*, 2006], pelo menos com divulgação visível para fora das paredes das Universidades e dos grandes centros de cálculo e computação.

Esta tentação, a que chamámos de popularidade do site, associado ao ranking (Secção 2.2.3), foi-se materializando, criando um outro gigante.

Fruto da ganância pelo dinheiro, o mesmo criador da 'Bela' criou o 'Monstro'.

Este 'Monstro', a que temos vindo a chamar de Spam e no nosso caso específico de Web Spam, tem-se tornado mais forte nos últimos anos, à medida que mais operadores ousaram utilizá-lo como um meio para aumentar o tráfego, na esperança de que possa também vir a aumentar as receitas. Fruto também de uma atenção - crescente - da comunidade académica, o crescimento tem vindo a ser refreado.

> A própria definição de 'web spam', que tem recebido com o tempo nuances sintácticas, aponta sempre para o que Ntoulas [Ntoulas *et al.*, 2006] define como: '*a injecção de páginas criadas artificialmente para influenciar os resultados dos motores de pesquisa, para orientar o tráfego de certas páginas para obtenção de lucros, ou simplesmente por prazer.*'

Evidenciámos que '*Qualquer que seja a definição é certo que Spam se refere a algo indesejável, mesmo perturbador, que influencia negati-*

vamente o processo de selecção de informação tratada em ambiente web, com utilização dos protocolos aí disponibilizados' (Secção 2.1.1).

Efectuado um levantamento dos diversos modos base de ataque do Web Spam, estudámos algumas das suas variantes, e bem assim algumas soluções encontradas para ultrapassar esses actos maldosos e repor os conceitos de credibilidade e de fiabilidade que, à margem estes inconvenientes, representam os sistemas de recuperação de informação. No caso da Web, os crawlers, os spiders, os motores de pesquisa, etc, são tudo elementos estruturais da ideia de sistemas de recuperação de informação que são alvos preferenciais.

A evolução para o maior aproveitamento das automatizações dos sistemas, como são as técnicas de 'Machine Learning', ou a adopção de algoritmos matemáticos mais precisos e complexos (ex. graph isomorphism) [Bharat *et al.*, 2001; Metaxas & DeStefano, 2005], estão cada vez mais inseridas nos processos de detecção de spam [Hidalgo, 2002; Sahami *et al.*, 1998]. Estas técnicas, no entanto, deparam-se com um forte constrangimento relacionado com a diferente disponibilização de conteúdos para os crawlers e para os browsers.

Em face dos muitos estudos que pretendem detectar e minimizar o spam, verificámos que há convergência no sentido de se agruparem em três grandes espécies: 'Link Spam', 'Content Spam' e 'Cloaking': As duas primeiras ligadas a técnicas de boosting do PR e a última a técnicas de camuflagem. É claro que esta é uma resposta à importância que é dada a esses elementos pelos principais algoritmos de Ranking [Brin & Page, 1998; Kleinberg, 1999; Metaxas & DeStefano, 2005].

Verificámos que, pelo número significativo de páginas infectadas - pelo menos 8% de todas as páginas indexadas são spam [Fetterly *et al.*, 2004; Metaxas & DeStefano, 2005] - , o Web Spam é um dos maiores desafios colocados a quem se dedica a recuperaçãode informação em ambiente web [Henzinger *et al.*, 2002; Metaxas & DeStefano, 2005]. Os estudos para limitar esses constrangimentos encontram dificuldades na identificação automática de spam com base apenas em algoritmos matemáticos [Bharat *et al.*, 2001; Metaxas & DeStefano, 2005]. Com efeito precisamos de compreender socialmente a questão do web spam e só depois analisar as questões técnicas, dado que é na área dos comportamentos sociais que o spam está sempre mais actualizado.

Também aqui se aplica, ainda que de forma adaptada, a máxima

de que pessoas boas se relacionam com pessoas boas, neste particular [Benczúr *et al.*, 2007b] spam normalmente aponta para spam e páginas boas apontam para páginas boas.

O futuro prevê um crescimento de complexidade dos ataques [Plc, 2008], pelo que qualquer especulação preditiva do que poderá vir a acontecer é utópica. Basta que nos debrucemos sobre a velocidade com que diariamente são detectados novos 'malwares', comparado com o que acontecia há cinco anos atrás, para nos apercebermos da seriedade da situação.

No entanto de algumas coisas temos a certeza:

O número e a variedade dos ataques continua a crescer, de forma cada vez mais estruturada em cadeias de crime organizado, que pretende usurpar informação e recursos.

A fuga de informação irá tornar-se cada vez mais preocupante, especialmente com a utilização crescente das tecnologias móveis. Muitos países já introduziram leis estritas sobre divulgação de informação. Estas leis visam punir todas as empresas que vasculham hipóteses para ultrapassar as barreiras de segurança, uma vez que, mesmo uma violação muito restrita de dados, uma vez divulgada, pode afectar a confiança global numa organização de produtos e serviços.

Também os Weblogs, ou simplesmente blogs, têm vindo a crescer como uma nova e importante forma de publicar informações, participar de discussões e formar comunidades. A crescente popularidade dos blogs tem dado origem a motores de pesquisa e análise centrada na 'blogosfera'.

Um requisito fundamental desses sistemas é o de identificar os blogs enquanto rastreiam a Web. Enquanto isso garante que somente os blogs serão indexados, os motores de busca são também muitas vezes sobrecarregado por 'splogs' (spam blogs), que acabam por influenciar negativamente as indexações. Embora de uma geração mais recente, esta é uma forma de spamdexing, que podemos facilmente incluir no grupo de 'content spam'.

A insegurança da web, fruto da sua forma concepcional, enfraquecida contra ataques remotos automatizados, fruto também dos crescentes modelos baseados em P2P, continuará a ser a principal forma de distribuição de malware específico.

É também neste sentido de especialização nesta nova área que uma parte da comunidade científica está a trabalhar. Kolaris [Kolari *et al.*, 2006b], por exemplo, usa um modelo identificado como SVM (Support Vector Machines) para identificação de blogs em que abre portas para desenvolvimentos desta tecnologia.

Cada um de nós, normais utilizadores de computadores, iremos sendo continuamente desafiados em questões de segurança e controlo dos nossos próprios equipamentos, para defesa a ataques criminosos, muitas vezes apenas pelo gosto de 'furarem' sistemas.

No entanto, quando tratados convenientemente, os problemas são sempre superáveis: melhorando as nossas defesas básicas, protecções actualizadas e um empenho pessoal de nos mantermos informados, podem proporcionar estabilidade nos nossos sistemas particulares e colectivos.

Em complemento, e como boas notícias, os softwares de segurança estão a melhorar dia-a-dia, inclusivamente produzindo alertas em defesa de possíveis novas formas de ataques.

Porque a guerra contra os motores de pesquisa continua a evoluir com rapidez, novas técnicas spam aparecerão, o que implica que novas abordagens anti-spam também serão desenvolvidas.

A luta vai continuar!

6.1 Futuro spam

A maioria dos trabalhos publicados discute métodos para combater spam de que já se conhecem atributos e formas de actuação. De facto raros são os trabalhos que abordam a questão de futuros ataques de spam. Mas, e porque a recuperação de informação (IR) na web é um campo com cada vez mais serviços de pesquisa, torna-se útil prever possíveis técnicas de spam que possam vir a aparecer, no sentido de se poder concretizar alguma antecipação.

Decididamente os algoritmos de ranking serão sempre o primeiro alvo. Daí que os spammers continuarão a atacar os factores a que esses algoritmos derem relevância. A pontuação dos modelos TF-IDF e de análise de links deverá continuar a ser utilizada pelos motores de pesquisa para elaborarem os rankings. Por isso os spammers continuarão a investir

em técnicas de manipulação desses valores, aumentando a complexidade dos métodos. Por exemplo no mesmo sentido de obter melhores valores na análise de links, eles podem desenvolver estruturas de links mais complexas e mais difíceis de detectar, ou, por outro lado, podem encontrar melhores técnicas de camuflagem, tais como o aprofundamento de Javascripts, para manipular esses valores.

Para além disso, se os motores de pesquisa anunciarem publicamente os novos componentes que venham a incluir nos novos algoritmos de ranking, serão, seguramente, alvos dos spammers. Como resultado, novas técnicas de spam irão aparecer.

A construção de sistemas de buscas patrocinadas, que apresentam publicidade na página de resposta dos motores de pesquisa[1], tem permitido grande sucesso financeiro.

Como esses sistemas estão agilizados para a realização de dinheiro, será provavel que esta seja uma área de grande investimento dos spammers. De facto algumas técnicas de spam dirigidas a esses sistemas foram já descobertas.

Por exemplo fraudes provenientes de clicks, que ocorrem nos sistemas de 'pay per click' quando uma pessoa, scripts automatizados, ou mesmo programas de computador, usam os browsers para clicarem em zonas de publicidade que geram pagamentos, sem terem qualquer interesse nos produtos para onde o link aponta. De facto, por força da componente financeira que representam, essas pesquisas patrocinadas podem ser novos focos de spam.

Os motores de pesquisa fornecem cada vez mais serviços relacionados com a recuperação de informação. Qualquer dos novos serviços que num futuro venha a ser disponibilizado será, seguramente, imediatamente estudado pelos spammers no sentido de encontrarem alguma forma de beneficiarem (de forma ilícita) desses novos serviços.

São já referências desta certeza, por exemplo, o spam sobre video, mas se a evolução da procura por pesquisas em Mapas se verificar então aí também virá a haver spam.

O 'spam baseado em imagem' tem se tornado uma técnica

[1]Exemplo: Google AdWords, Google Adsense, Yahoo! Search Marketing, Microsoft Adcenter

popular entre os propagadores de spam devido à sua capacidade de ignorar tecnologias tradicionais de filtragem anti-spam. Ao invés de enviar mensagens como texto com ou sem imagens anexadas, os propagadores de spam passaram a enviar mensagens que contêm somente imagens.

O spam de imagem - mais comum - é uma mensagem de e-mail não solicitada que contém somente uma imagem (normalmente um arquivo incorporado no formato .JPG ou .GIF). Essa imagem é formatada para conter a mensagem que o propagador do spam deseja transmitir. Pode ser que o e-mail contenha uma figura ou algum 'texto', porém esse 'texto' faz parte da imagem. Os propagadores de spam também tentam confundir os filtros variando subtil e levemente as imagens em cada e-mail. Essas alterações não são visíveis (e são irrelevantes para o leitor), mas tornam muito difícil para as tecnologias anti-spam detectá-las como um ataque de spam único, uma vez que todas as 'assinaturas' do spam são diferentes.

O spam de imagem tem mostrado um crescimento explosivo recentemente. Richi Jennings, analista da Ferris Research, afirma que o número de e-mails de spam de imagem aumentou aproximadamente 900% durante o último ano. O spam de imagem consome também uma grande quantidade de largura de banda e espaço de armazenamento. [...]

Também no Twiter:

O Twitter tem um problema sério de spam. Qualquer utilizador que seja um pouco mais conhecido, por ter um blog, ou apenas por divulgar seu username em redes sociais ou outros serviços da web, tem sido convidado por 'pessoas' no mínimo curiosas.

A maior parte dos novos seguidores estranhos do seu Twitter são de empresas disfarçadas de utilizadores, que esperam que adopte a pseudo-etiqueta do serviço e os siga também, passando a receber os seus updates publicitários. Se por um lado o internauta mais entendido vai perceber que o 'iphone4free' não pode ser coisa boa, a chegada dos spammers cria resis-

> tência na checagem dos novos assinantes, que pode render belo conteúdo.
>
> A solução imediata para afastar os spammers é a mesma adotada em outros serviços - o bom e velho 'captcha' junto com uma descrição textual do motivo para assinar o conteúdo do outro. Acaba a transparência do processo, mas evita-se a infestação do spam.[2]

Não importa a complexidade das técnicas de spam uma vez que todas elas são projectadas para manipular os factores determinantes nos algoritmos de classificação. Por exemplo, porque os spammers continuam a acreditar que os algoritmos baseados em estruturas de links continuarão a ser usados, continuarão a investir tempo e dinheiro tentando manipular essas estruturas.

Embora o futuro spam possa assumir diferentes formas e métodos, podemos combatê-los incidindo sobre as características que têm permanecido relativamente inalteradas ou utilizando algumas metodologias anti-spam já testadas e amadurecidas.

De forma a obter uma melhor classificação para as suas páginas, os spammers têm necessidade de reforçar alguns pontos, o que fará com que as páginas manipuladas apresentem algumas características especiais, comparadas com páginas normais [Benczúr *et al.*, 2005; Fetterly *et al.*, 2004]. Por isso, a detecção dessas características especiais é um bom indicador para a presença de Spam. Por exemplo, sites com muitos links incoming e outgoing iguais podem ser presença de um link farm. Do mesmo modo, a inserção de grande quantidade de palavras (eventualmente palavras-chave) [Davison, 2000; Drost & Scheffer, 2005], ainda que soltas, ou o povoamento das mesmas páginas por links incaracterísticos são provavelmente fruto de preenchimento pouco sério. Neste sentido, temos de procurar mais recursos e páginas com características especiais. Por exemplo, se a maioria dos outgoing links de um site apontam para outros sites que proporcionam técnicas de 'affiliate' então este site usa técnicas 'affiliate'.

Associada a esta abordagem estatística de contagem e análise de links, os avanços nas técnicas de 'machine learning' podem inovar procedimentos nomeadamente com a possibilidade de abordagem diferente

[2] http://futuro.vc/tag/spam/

de um mesmo algoritmo em face de algumas classes de características que sejam detectadas. Por exemplo, pode construir-se um classificador (Secção 4.2.1 [Ntoulas *et al.*, 2006]) para saber se um clique é um clique fraudulento, ou podemos utilizar um classificador para saber se um link é um link spam ou não, entre outras imensas possibilidades. É cada vez mais claro que há misturas de técnicas baseadas em links, com outras técnicas baseadas nos conteúdos [Gibson *et al.*, 2005].

O 'braço de ferro' entre spammers e motores de pesquisa é já longo. Os contra-ataques dos spammers depois de os motores de pesquisa neutralizarem prévios ataques, mantém actual esta designação. O trunfo que entretanto os investigadores começam a usar nas novas técnicas de spam, ligadas à preemptividade dos malefícios, pode desmotivar os 'hackers' durante, pelo menos, algum tempo.

Quanto maior for a disponibilidade de colaboração, quer do mundo industrial quer do mundo científico, então melhores e mais sofisticados algoritmos anti-spam poderão ser produzidos, por um lado, e por outro, quanto mais evoluirmos socialmente, no sentido de minimizarmos os constrangimentos encontrados na necessária unanimidade na classificação humana dos servidores, menos credibilidade daremos a impostores.

Serão factores preponderantes para a diminuição do investimento em spam quando:

- For mais caro produzir páginas de spam do que páginas oficiais com qualidade, ou
- For mais caro produzir páginas de spam do que os lucros que daí advenham.

Pela mesma razão de, como diz o povo, o óptimo é inimigo do bom, também a vitória não requer obrigatoriamente a perfeição. Por isso temos esperança de que as pesquisas contínuas possam tornar a 'pirataria mais cara do que o original', ou seja, de que o genuíno compensa.

Assim sendo, sairemos vencedores!

Bibliografia

Abernethy, Jacob, Chapelle, Olivier, & Castillo, Carlos. 2008a. Web spam Identification Through Content and Hyperlinks. *AIRWeb '08, Beijing, China*, May, 22.

Abernethy, Jacob, Chapelle, Olivier, & Castillo, Carlos. 2008b. *WITCH: A New Approach to Web Spam Detection.* Tech. rept. Yahoo! Research.

Acharya, A., Cutts, M., Dean, J., Haahr, P., Henzinger, M., Hoelzle, U., Lawrence, S., Pfleger, K., Sercinoglu, O., & Tong, S. 2005. *Information Retrieval based on Historical Data.*

Adali, Sibel, Liu, Tina, & Magdon-Ismail, Malik. 2005 (May). Optimal Link Bombs are Uncoordinated. *In: Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web.*

Agichtein, Eugene, Brill, Eric, & Dumais, Susan. 2006. Improving Web Search Ranking by Incorporating User Behavior Information. *SIGIR ´06, Seattle, Washington, USA*, August 6-11.

Alliance, Cyber Security Industry. 2007 (January). *SPAM: Get the Facts.* Tech. rept. Cyber Security.

Alves, Manuel. 2008. Criação de Modelos de Língua de suporte a Sistemas de Tradução. Departamento de Informática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal.

Attenberg, Josh, & Suel, Torsten. 2008. Cleaning search results using term distance features. *Pages 21–24 of: AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM.

Attia, Elham. 2006. Web Spam. *Institut für Informationsverarbeitung und Prozessmanagement Abteilung für Informationswirtschaft*, Jun,26.

Baeza-Yates, R., Castillo, C., Junqueira, F., Plachouras, V., & Silvestri, F. 2007a. Challenges on distributed web retrieval. *ICDE*, 6–20.

Baeza-Yates, Ricardo, & Castillo, Carlos. 2007. Crawling the Infinite Web. *Journal of Web Engineering*, **6**(1), 49–72.

Baeza-Yates, Ricardo, & Ribeiro-Neto, Berthier. 1999. *Modern Information Retrieval*. Addison Wesley.

Baeza-Yates, Ricardo, Castillo, Carlos, & López, Vicente. 2005 (May-10). Pagerank Increase under Different Collusion Topologies. *In:* Davison, Brian D. (ed), *Proceedings of First International Workshop on Adversarial Information Retrieval on the Web*.

Baeza-Yates, Ricardo, Castillo, Carlos, Junqueira, Flavio, Plachouras, Vassilis, & Silvestri, Fabrizio. 2007b. Challenges in Distributed Information Retrieval. *In: International Conference on Data Engineering (ICDE)*. Istanbul, Turkey: IEEE CS Press.

Baeza-Yates, Ricardo, Castillo, Carlos, & Efthimiadis, Efthimis. 2007c. Characterization of national Web domains. *ACM Transactions on Internet Technology*, **7**(2).

Becchetti, Luca, Castillo, Carlos, Donato, Debora, & Fazzone, Adriano. 2006a (August). A comparison of sampling techniques for Web characterization. *In: Workshop on Link Analysis (LinkKDD)*.

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Leonardi, Stefano, & Baeza-Yates, Ricardo A. 2006b. Link-based characterization and detection of web spam. *In: In AIRWeb*.

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Leonardi, Stefano, & Baeza-Yates, Ricardo. 2006c. Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection. *In: Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD)*. Pennsylvania, USA: ACM Press.

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Leonardi, Stefano, & Baeza-Yates, Ricardo. 2008. Web Spam Detection: Link-based and Content-based Techniques. *Pages 99–113 of:* Friedhelm (ed), *The European Integrated Project Dynamically Evolving, Large Scale Information Systems (DELIS): proceedings of the final workshop*, vol. 222. Heinz-Nixdorf-Institut, Universität Paderborn.

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Baeza-Yates, Ricardo, & Stefano, Leonadi. to appear (In Press). Link Analysis for Web Spam Detection. *ACM Transactions on the Web Journal(TWJ)*, March. Work in Progress.

Benczúr, András, Bíró, István, Csalogány, Károly, & Sarlós, Tamás. 2007a. Web Spam Detection via Commercial Intent Analysis. *AIRWeb, 2007 Banff, Alberta, Canada*, May 8.

Benczúr, András A., Csalogány, Károly, Sarlós, Tamás, & Uher, Máté. 2005. Spamrank - fully automatic link spam detection. *In: In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb).*

Benczúr, András A., Biró, I., Csalogány, Károly, & Uher, Károly. 2006a. *Detecting nepotistic links by language model disagreement.*

Benczúr, András A., Csalogány, Károly, & Sarlós, Tamás. 2006b. Link-Based Similarity Search to Fight Web Spam. *Pages 9–16 of: AIRWeb.*

Benczúr, András A., Csalogány, Károly, Lukács, László, & Siklósi, Dávid. 2007b. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn?

Benevenuto, Fabricio, Rodrigues, Tiago, Almeida, Virgilio, Almeida, Jussara, Zhang, Chao, & Ross, Keith. 2008. Identifying Video Spammers in Online Social Networks. *ACM 978-1-60558-159-0.*

Berners-Lee, T., Fielding, R., & Masinter, L. 2005 (January). RFC 3986: Uniform Resource Identifier (URI): Generic Syntax.

Bharat, K., Chang, B. W., Henzinger, M. R., & Ruhl, M. 2001. Who links to whom: Mining linkage between web sites. *IEEE Computer Society*, **Proceedings of the 2001 IEEE International Conference on Data Mining**, 51 – 58.

Bharat, Kirshna, & Mihaila, George A. 2000. Hilltop: A search engine based on expert documents. In Poster proceedings of WWW9, pages 72-73, 2000. *In Poster proceedings of WWW*, 72–73.

Bharat, Krishna, & Henzinger, Monika R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. *Pages 104–111 of: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* New York, NY, USA: ACM.

Bharat, Krishna, & Mihaila, George A. 2002. When experts agree: using non-affiliated experts to rank popular topics. *ACM Trans. Inf. Syst.*, **20**(1), 47–58.

Bifet, Albert, Castillo, Carlos, Chirita, Paul A., & Weber, Ingmar. 2005. An Analysis of Factors Used in a Search Engine's Ranking. *In: First International Workshop on Adversarial Information Retrieval on the Web.*

Boldi, P., & Vigna, S. 2003. *The WebGraph framework I: Compression techniques.*

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. 2002. *Ubicrawler: A scalable fully distributed web crawler.*

Bordignon, Fernando R. A., & Tolosa, Gabriel H. 2007. Recuperación de Información:Un área de investigación en crecimiento. *Revista Electronica de Estudios Telemáticos*, **6(1)**, 53 – 76.

Borodin, Alan, Roberts, Gareth O., Rosenthal, Jeffrey S., & Tsaparas, Panayiotis. 2001. Finding authorities and hubs from link structures on the World Wide Web. *Pages 415–429 of: World Wide Web.*

Borodin, Allan, Roberts, Gareth O., Rosenthal, Jeffrey S., & Tsaparas, Panayiotis. 2005. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, **5**(1), 231–297.

Bíró, István, Szabó, Jácint, & Benczur, András A. 2008. Latent Dirichlet Allocation in Web Spam Filtering. *AIRWeb '08, Beijing, China*, May, 22.

Brin, S., & Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30(1 - 7)**, 107 – 117.

Broder, A. 2002. A taxonomy of web search. *SIGIR Forum*, **36(2)**, 3 – 10.

Buckley, Chris, & Salton, Gerard. 1995. Optimization of relevance feedback weights. *Pages 351 – 357 of: SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. Seattle, Washington, United States: ACM.

Buehrer, Greg, Stokes, Jack W., & Chellapilla, Kumar. 2008. A large-scale study of automated web search traffic. *Pages 1–8 of: AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Burdon, David. 2005. *The Basics of Search Engine Optimisation*. Simply Clicks.

Cafarella, Mike, & Cutting, Doug. 2004. Building Nutch: Open Source Search: A case study in writing an open source search engine. *ACM Queue*, **2**(2).

Caminero, Rizza Camus, & Mikami, Yoshiki. 2008. The Link Structure of Language Communities and its Implication for Language-specific Crawling. *The 6th Workshop on Asian Languae Resources*.

Castilho, Carlos, Donato, Debora, Becchetti, Luca, Boldi, Paolo, Leonardi, Stefano, Santini, Massimo, & Vigna, Sebastiano. 2006. A Reference Collection for Web Spam. *ACM SIGIR Forum*, **Volume 40 , Issue 2 (December 2006)**(December), 11–24.

Castillo, Carlos. 2004 (November). *Effective Web Crawling*. Ph.D. thesis, University of Chile.

Castillo, Carlos. 2008 (Jan, 15). *Resources for Research on Web Spam*.

Castillo, Carlos, Donato, Debora, Gionis, Aristides, Murdock, Vanessa, & Silvestri, Fabrizio. 2007a. Know your Neighbors: Web Spam Detection using the Web Topology. *In: Proceedings of SIGIR*. Amsterdam, Netherlands: ACM.

Castillo, Carlos, Chellapilla, Kumar, & Davison, Brian D. 2007b (May, 8). Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web. *In: AIRWeb 2007*.

Castillo, Carlos, Chellapilla, Kumar, & Denoyer, Ludovic. 2008a. *AIRWeb 2008 - Fourth International Workshop on Adversarial Information Retrieval on the Web*.

Castillo, Carlos, Corsi, Claudio, Donato, Debora, Ferragina, Paolo, & Gionis, Aristides. 2008b. Query-log mining for detecting spam. *Pages 17–20 of: AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Caverlee, James, Webb, Steve, & Liu, Ling. 2007. Spam-Resilient Web Rankings via Influence Throttling. *Pages 1–10 of: IPDPS*. IEEE.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., & Rajagopalan, S. 1998a. Automatic resource compilation by analyzing hyperlink structure and associated text. Proc. of the 7th World-Wide Web Conference (WWW7).

Chakrabarti, Soumen. 2001. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. *Pages 211–220 of: WWW '01: Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM.

Chakrabarti, Soumen, Dom, Byron E., Gibson, David, Kumar, Ravi, Raghavan, Prabhakar, Rajagopalan, Sridhar, & Tomkins, Andrew. 1998b. *Experiments in Topic Distillation*. Tech. rept. IBM Alamaden Research Center, San Jose, CA.

Chakrabarti, Soumen, Joshi, Mukul, & Tawde, Vivek. 2001. Enhanced topic distillation using text, markup tags, and hyperlinks. *Pages 208–216 of:* Kraft, Donald H., Croft, W. Bruce, Harper, David J., & Zobel, Justin (eds), *Proceedings of the 24th An-*

nual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press.

Chambers, R. 2006. *Search engine strategies: a model to improve website visibility for SMME websites*. M.Phil. thesis, Cape Peninsula University of Technology.

Chellapilla, Kumar, & Chickering, David Maxwell. 2006. Improving Cloaking Detection using Search Query Popularity and Monetizability. *Pages 17–23 of: AIRWeb*.

Chellapilla, Kumar, & Maykov, Alexey. 2007. A taxonomy of JavaScript redirection spam. *Pages 81–88 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Chen, Shuo, Cristofor, Laurentiu, Kiciman, Emre, & Kumar, Arunvijay. 2006 (December, 12). *Strider Search Defender: Automatic and Systematic Discovery of Search Spammers through Non-Content Analysis*. Tech. rept. MSR-TR-2006-97. Microsoft Research Strider Team.

Cho, J., Garcia-Molina, H., & Page, L. 1998. Efficient Crawling Through URL Ordering. *In: In Proceedings of 7th World Wide Web Conference*.

Collins, Gord. 2004 (August 9th). *Latest Search Engine Spam Techniques*.

Corey, T. S. 2001. Catching on-line traders in a web of lies: The perils of internet stock fraud. *Ford Marrin Esposito, Witmeyer & Glesser, LLP*, May.

Costa, Valéria O., & Fernandes, Juliana C. 2007. Uma análise do comportamento das máquinas de busca da web. *Centro Federal de Educação Tecnológica do Piauí - Cadernos Temáticos*, **15**(Mar), 37–42.

da Costa Carvalho, André Luiz, Chirita, Paul Alexandru, de Moura, Edleno Silva, Calado, Pável, & Nejdl, Wolfgang. 2006. Site level noise removal for search engines. *Pages 73–82 of: WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM.

Davison, Brian D. 2000. Recognizing Nepotistic Links on theWeb. *In: Workshop on Artificial Intelligence for web search.*

Davison, Brian D., Gerasoulis, Apostolos, Kleisouris, Konstantinos, Lu, Yingfang, ju Seo, Hyun, Wang, Wei, & Wu, Baohua. 1999a. DiscoWeb: Applying link analysis to web search. *Pages 148–149 of: In Poster Proceedings of the Eighth International World Wide Web Conference.*

Davison, Brian D., Gerasoulis, Apostolos, Kleisouris, Konstantinos, Lu, Yingfang, ju Seo, Hyun, Tian, Junyu, Wang, Song, Wang, Wei, & Wu, Baohua. 1999b. An Early DiscoWeb Prototype at TREC8. *In: TREC.*

Donato, Debora, Paniccia, Mario, Selis, Maddalena, Castillo, Carlos, Cortese, Giovanni, & Leonardi, Stefano. 2007 (May). *New Metrics for Reputation Management in P2P Networks.* Tech. rept. Banff, Alberta, Canada.

dos Santos Batista Junior, Wilson. 2006. *Recuperação de Informação com auxilio de extratos automáticos.* M.Phil. thesis, Universidade Federal de São Carlos. Brasil.

Drost, Isabel, & Scheffer, Tobias. 2005. Thwarting the nigritude ultramarine: learning to identify link spam. *Pages 233–243 of: Proceedings of the 16th European Conference on Machine Learning (ECML).* Lecture Notes in Artificial Intelligence, vol. 3720.

Du, Ye, Shi, Yaoyun, & Zhao, Xin. 2007. Using spam farm to boost PageRank. *Pages 29–36 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM.

Edwards, Jenny, McCurley, Kevin, & Tomlin, John. 2001. An Adaptive Model for Optimizing Performance of an Incremental Web Crawler. *WWW10, Hong Kong.*, May 1 - 5.

Eiron, Nadav, McCurley, Kevin S., & Tomlin, John A. 2004. Ranking the web frontier. *Pages 309–318 of: WWW '04: Proceedings of the 13th international conference on World Wide Web.* New York, NY, USA: ACM.

Fattahi, Rahmatollah, Wilson, Concepción S., & Cole, Fletcher. 2008. An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents. *Inf. Process. Manage.*, **44**(4), 1503–1516.

Fetterly, Dennis, Manasse, Mark, & Najork, Marc. 2004. Spam, Damn Spam, and Statistics. *Seventh International Workshop on the Web and Databases (WebDB 2004)*, June 17-18. Paris, France.

Gan, Qingqing, & Suel, Torsten. 2007. Improving web spam classifiers using link structure. *Pages 17–20 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM.

Geng, Guanggang, Wang, Chunheng, & Li, Qiudan. 2008. Improving web spam detection with re-extracted features. *Pages 1119–1120 of: WWW '08: Proceeding of the 17th international conference on World Wide Web.* New York, NY, USA: ACM.

Gerasoulis, Apostolos, Wang, Wei, & Seo, Hyun-Ju. 2006. *Retrieval and display of data objects using a cross-group ranking metric.*

Gibson, David, Kumar, Ravi, & Tomkins, Andrew. 2005. Discovering large dense subgraphs in massive graphs. *Pages 721–732 of: VLDB '05: Proceedings of the 31st international conference on Very large data bases.* VLDB Endowment.

Google. 2008. *Cloaking, sneaky Javascript redirects, and doorway pages.*

Graham, L., & Metaxas, P. T. 2003. Of course it´s true; I saw it on the internet! Critical thinking in the internet era. *Commun. ACM*, **46(5)**, 70 – 75.

Guha, R., Kumar, R., Raghavan, P., & Tomkins, A. 2004. Propagation of Trust and Distrust. *In: International World Wide Web Conference.*

Gupta, Atul. 2003. Analysis and Implications of Hilltop Algorithm. December.

Gyongy, Zoltán, Garcia-Molina, Hector, & Pedersen, Jan. 2004. Combating Web Spam with TrustRank. *Proceedings of the 30th*

VLDB Conference, Toronto, Canada. Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

Gyongyi, Z., & Garcia-Molina, H. 2005. Web spam taxonomy. *First International Workshop on Adversarial Information Retrieval on the Web*.

Gyöngyi, Zoltán, & Garcia-Molina, Hector. 2005. Link spam alliances. *Pages 517–528 of: VLDB '05: Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment.

Gyongyi, Zoltan, Berkhin, Pavel, Garcia-Molina, Hector, & Pedersen, Jan. 2006. Link spam detection based on mass estimation. *Pages 439–450 of: VLDB '06: Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment.

Haveliwala, Taher, & Kamvar, Sepandar. 2003. *The Second Eigenvalue of the Google Matrix*. Tech. rept. 20. Stanford University.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. *Pages 9–16 of: Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics.

Hearst, Marti A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.

Henzinger, M. R. 2001. Hyperlink analysis for the web. *IEEE Internet Computing*, **5(1)**, 45 – 50.

Henzinger, M. R., Motwani, R., & Silverstein, C. 2002. Challenges in web search engines. *SIGIR Forum*, **36(2)**, 11 – 22.

Henzinger, Monika R. 2002 (December). Algorithmic Challenges in Web Search Engines. vol. 1.

Hidalgo, José María Gómez. 2002. Evaluating cost-sensitive Unsolicited Bulk Email categorization. *Pages 615–620 of: SAC '02: Proceedings of the 2002 ACM symposium on Applied computing*. New York, NY, USA: ACM.

Hiemstra, Djoerd, & Robertson, Stephen. 2001. Relevance Feedback for Best Match Term Weighting Algorithms in Information

Retrieval. *In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries.*

Hilberer, Michael, & Speck, Hendrick. 2005. Development of algorithms for web spam detection based on structure and link analysis. *IADIS INTERNATIONAL JOURNAL ON WWW/INTERNET (ISSN: 1645-7641)*, **Vol.3 Issue 2**.

Hindman, M., Tsioutsiouliklis, K., & Johnson, J. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. *Annual Meeting of the Midwest Political Science Association*, April.

Hoeschl, Hugo Cesar. 2006 (February). Revista Consultor Jurídico. *In: Tecnologia online - Ontoweb: A nova era das ferramentas de busca.*

Jansen, Bernard J., Spink, Amanda, & Saracevic, Tefko. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, **36**(2), 207–227.

Jiang, Qiancheng, Zhang, Lei, Zhu, Yizhen, & Zhang, Yan. 2008. Larger is better: seed selection in link-based anti-spamming algorithms. *Pages 1065–1066 of: WWW'08: Proceeding of the 17th international conference on World Wide Web.* New York, NY, USA: ACM.

Johnson, Steven. 2003. Emergência: a dinâmica de redes em formigas, cérebros, cidades e softwares. *Jorge Zahar : Rio de Janeiro*, 94.

Kg, N. 1999. A maximum likelihood ratio information retrieval model. *Pages 214–221 of: Proceedings of the 8th Text Retrieval Conference TREC-8.*

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46(5)*, 604 – 632.

Kobayashi, Mei, & Takeda, Koichi. 2000. Information retrieval on the web. *ACM Computing Surveys*, **32**(2), 144–173.

Kolari, Pranam, Java, Akshay, Finin, Tim, Oates, Tim, & Joshi, Anupam. 2006a. Detecting Spam Blogs: A Machine Learning Approach. *In: AAAI.*

Kolari, Pranam, Finin, Tim, & Joshi, Anupam. 2006b. SVMs for the Blogosphere: Blog Identification and Splog Detection. *In: AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*University of Maryland, Baltimore County, for Computer Science and Electrical Engineering.

Koutrika, Georgia, Effendi, Frans A., Gyöngyi, Zoltán, Heymann, Paul, & Garcia-Molina, Hector. 2007. Combating spam in tagging systems. *Pages 57–64 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM Press.

Krause, Beate, Schmidt, Christoph, Hotho, Andreas, & Stumme, Gerd. 2008. The Anti-Social Tagger - Detecting Spam in Social Bookmarking Systems. *AIRWeb '08*, May, 22.

Krishnan, Vijay, & Raj, Rashmi. 2006. Web Spam Detection with Anti-Trust Rank. *Pages 37–40 of: AIRWeb.*

Kumar, Ravi, Raghavan, Prabhakar, Rajagopalan, Sridhar, Sivakumar, D., Tomkins, Andrew S., & Upfal, Eli. 2000. *The Web as a graph.*

Lancaster, F. W. 1993. *Indexação e resumos: teoria e prática.* Brasília : Briquet de Lemos/Livros.

Langville, Amy N., & Meyer, Carl D. 2004. Deeper inside PageRank. *Internet Mathematics*, **1**(3), 335–380.

Lee, A. M., & Lee(eds.), E. B. 1939. *The Fine Art of Propaganda.* The Institute for Propaganda Analysis.

Lempel, R., & Moran, S. 2001. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, **19**(2), 131–160.

Levene, Mark, & Poulovassilis, Alexandra. 2004. *Web Dynamics - Adapting to Change in Content, Size, Topology and Use.* Springer.

Li, Longzhuang, Shang, Yi, & Zhang, Wei. 2002. Improvement of HITS-based algorithms on web documents. *Pages 527–535 of: WWW '02: Proceedings of the 11th international conference on World Wide Web.* New York, NY, USA: ACM Press.

Lin, Jun-Lin. 2009. Detection of cloaked web spam by using tag-based methods. *Expert Syst. Appl.*, **36**(4), 7493–7499.

Lin, Yu-Ru, Sundaram, Hari, Chi, Yun, Tatemura, Junichi, & Tseng, Belle L. 2007. Splog detection using self-similarity analysis on blog temporal dynamics. *Pages 1–8 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Liu, Yiqun, Cen, Rongwei, Zhang, Min, Ma, Shaoping, & Ru, Liyin. 2008. Identifying Web Spam with User Behavior Analysis. *AIRWeb '08, Beijing, China*, April, 22.

Lévy, Pierre. 1999. *Cibercultura*. São Paulo: Editora.

Lynch, C. A. 2001. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *J. Am. Soc. Inf. Sci. Technol*, **52(1)**, 12 – 17.

M. Bianchini, M. Gori, & Scarselli., F. 2003 (Oct). PageRank and web communities. Web Intelligence Conference.

Manning, C. D., & Schutze, H. 2001. *Foundations of statistical natural language processing*. MIT Press.

Manning, Christopher D., & Schtze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Marchiori, M. 1997. The quest for correct information on the web: hyper search engines. *Comput. Netw. ISDN Syst.*, **29(8-13)**, 1225 – 1235.

Marcondes, Carlos Henriques. 2007. Metadados: descrição e recuperação de informação na Web. May.

Marcondes, Carlos Henriques, & Sayão, Luis Fernando. 2002. Documentos digitais e novas formas de cooperação entre sistemas de informação. *C&t. Ciência da Informação, Brasília*, **31**(3), 42 – 54.

Martins, Francisco Menezes, & da Silva, Juremir Machado. 2004. *A genealogia do virtual*. Isbn-10: 8520503470 - isbn-13: 9788520503478 edn. Sulina.

Mbikiwa, F., & Weideman, M. 2006. *Implications of search engine spam on the visibility of South African e-commerce Web sites.* ISSN: 1560683x, Volume 8, Issue 4.

McGaffin, Ken. 2005. *The Financial Times Website and Hidden Links.* visitado em 2009-03-12.

Mcnichol, Tom. 2004. Engineering Google Results to Make a Point. *nytimes.com*, Jan, 22.

Metaxas, Panagiotis T., & DeStefano, Joseph. 2005. Web Spam, Propaganda and Trust. *AIRWeb2005, May 10, 2005*, 5.

Miller, D. R. H., Leek, T., & Schwartz, R. M. 1999. A hidden Markov model information retrieval system. *Proceedings of the 22nd ACM SIGIR Conference*, 214–221.

Mishne, Gilad. 2005. Blocking Blog Spam with Language Model Disagreement. *In: In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb.*

Najork, M. 2003. *System and method for identifying cloaked web servers.*

Ng, Andrew Y., Zheng, Alice X., & Jordan, Michael I. 2001a. Link Analysis, Eigenvectors and Stability. *Pages 903–910 of: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).*

Ng, Andrew Y., Zheng, Alice X., & Jordan, Michael I. 2001b. Stable algorithms for link analysis. *Pages 258–266 of: SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* New York, NY, USA: ACM.

Nie, Lan, Wu, Baoning, & Davison, Brian D. 2007 (Jan). *Incorporating Trust into Web Search.* Dept. of Computer Science and Engineering - Lehigh University, Bethlehem, PA, 1801.

Ntoulas, Alexandros, Najork, Marc, Manasse, Mark, & Fetterly, Dennis. 2006. Detecting Spam Web Pages through Content Analysis. *International World Wide Web Conference Committee (IW3C2)*, May 23-26.

Oliveira, Walter Clayton, & Vidotti, Silvana Aparecida Borsetti Gregorio. 2004. Auto-organização do ciberespaço: Uma visão holística. *Textos de la CiberSociedad*, **4**.

Pant, G., Srinivasan, P., & Menczer, F. 2004. Crawling the Web.

Parreira, Josiane X., Donato, Debora, Castillo, Carlos, & Weikum, Gerhard. 2007. Computing Trusted Authority Scores in Peer-to-Peer Web Search Networks. *In: AIRWeb*. ACM International Conference Proceeding Series, vol. 215.

Perkins, Alan. 2001 (September, 30). *The classification of Search Engine Spam*. Tech. rept.

Peters, C., Gey, F., Gonzalo, J., Mueller, H., Jones, G.J.F., Kluck, M., Magnini, B., & de Rijke, M. (eds). 2006. *Use of Free On-line Machine Translation for Interactive Cross-Language Question Answering*. Lecture Notes in Computer Science, vol. 4022. Springer.

Plc, Sophos. 2008 (December). *Security threat report: 2009*. Tech. rept. Sophos.

Ponte, J. M., & Croft, W. B. 1998. A language modelling approach to IR. *Proceedings of th 21st ACM SIGIR Conference*, 275–281.

Pringle, G., Allison, L., & Dowe, D. L. 1998. What is a tall poppy among web pages? *Elsevier Science Publishers B. V.,Proceedings of the seventh international conference on World Wide Web 1998.*, **7**, 369–377.

Qi, Xiaoguang, Nie, Lan, & Davison, Brian D. 2007. Measuring similarity to detect qualified links. *Pages 49–56 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Raghavan, Prabhakar, Rajagopalan, Sridhar, Sivakumar, D, Tomkins, Andrew, & Upfal, Eli. 2001. *Stochastic models for the web graph Ravi Kumar*.

Richardson, Matthew, Prakash, Amit, & Brill, Eric. 2006. Beyond PageRank: Machine Learning for Static Ranking. *World Wide Web Conference Committee (IW3C2) WWW 2006, Edinburgh, Scotland.*, May 23-26.

Risvik, Knut Magne, & Michelsen, Rolf. 2002. Search Engines and Web Dynamics. Computer Networks, vol. 39.

Roberts, Gareth O., & Rosenthal, Jeffrey S. 2003. Downweighting tightly knit communities in world wide web rankings. *Page 2003 of: Advances and Applications in Statistics (ADAS*. AddisonWesley.

Robertson, Stephen E., & Jones, Karen Sparck. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, **27**(May-Jun), 129–146.

Sahami, Mehran, Dumais, Susan, Heckerman, David, & Horvitz, Eric. 1998 (July). A Bayesian Approach to Filtering Junk E-mail. *In: AAAI Workshop on Learning for Text Categorization*.

Saito, Hiroo, Toyoda, Masashi, Kitsuregawa, Masaru, & Aihara, Kazuyuki. 2007. A large-scale study of link spam detection by graph algorithms. *Pages 45–48 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

Salton, Gerard, & McGill, Michael. 1984. Introduction to Modern Information Retrieval.

Schons, Claudio Henrique. 2007. El volumen de la información en el Internet y su desorganización: reflexiones e perspectivas. *Informação & Informação, Londrina*, **12 (1)**(Jan./Jun.).

Shen, Guoyang, Gao, Bin, Liu, Tie-Yan, Feng, Guang, Song, Shiji, & Li, Hang. 2006. Detecting Link Spam Using Temporal Information. *Pages 1049–1053 of: ICDM '06: Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, **33(1)**, 6 – 12.

Smigler, Scott. 2005. An ethical alternative to "Doorway Pages". *http://www.webpronews.com/insiderreports/2005/02/16/an-ethical-alternative-to-doorway-pages*, Feb, 16.

Sánchez, Montserrat Mateos. 2006 (07). *Aplicación de Técnicas de Clustering en la Recuperación de Información Web*. Ph.D. thesis, Universidad de Salamanca - Departamento de Informática y Automática. Director Dr. D. Carlos García-Figuerola Paniagua.

Sobek, M. 2002. PR0 - Google's PageRank 0 penalty.

Souza, Renato Rocha, & Alvarenga, Lídia. 2004. A web semântica e suas contribuições para a Ciência da Informação. *Ciência da Informação*, **33(1)**(Jan./Abr.), 132 – 141.

Sparck-Jones, K., Walker, S., & Robertson, S.E. 2000. A probabilistic model of information retrieval: development and comparative experiments. *In: Information Processing & Management 36(6), pp. 779-840.*

Svore, Krysta M., Vanderwende, Lucy, & Burges, Christopher J.C. 2007a. Enhancing Single document Summarization by Combining RankNet and Third - party Sources.

Svore, Krysta M., Wu, Qiang, Burges, Chris J. C., & Raman, Aaswath. 2007b. Improving web spam classification using rank-time features. *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, **215**, 9–16. ISBN:978-1-59593-732-2.

Sydow, Marcin, Piskorski, JAcub, Weiss, Dawid, & Castillo, Carlos. 2008. Application of Machine Learning in Combating Web Spam. Para publicação.

Taveira, Danilo Michalczuk, Moraes, Igor Monteiro, Rubinstein, Marcelo Gonçalves, & Duarte, Otto Carlos Muniz Bandeira. 2006. Técnicas de Defesa Contra Spam. *SBSeg ´06: Livro Texto dos Minicursos*, August, 202–250.

Thurow, Shari. 2003. Search Engine Spam. Dec 8,. ClickZ.

Tolosa, Gabriel, Bordignon, Fernando, Baeza-Yates, Ricardo, & Castillo, Carlos. 2007a. Characterization of the Argentinian Web. *Cybermetrics*, **11**(1), 3+.

Tolosa, Gabriel, Bordignon, Fernando, Baeza-Yates, Ricardo, & Castillo, Carlos. 2007b. Distinctive Features of the

Argentinian Web. *Proceedings of LA-WEB. Santiago, Chile, 2007. IEEE CS PressPress.*

Visser, E. B., Kritzinger, W. T., & Weideman, M. 2007 (6 - 8 September). *Search engine optimisation elements and their effect on Website visibility: implementation of the Chambers model.* Proceedings of the 8th Annual Conference on WWW Applications. South Africa, Bloemfontein.

Ward, Darrin. 2003. Google - What Is PR0 (PageRank Zero). July.

Webb, Steve. 2006. Introducing the Webb spam corpus: Using email spam to identify Web spam automatically. *In: In Proceedings of the 3rd Conference on Email and AntiSpam (CEAS) (Mountain View.*

WEBSPAM-UK2007, Team. 2007a. *Classification interface for WEBSPAM-UK2007.* http://www-connex.lip6.fr/ xmlmining/ws/assesment/cgi-bin/host.php.

WEBSPAM-UK2007, Team. 2007b (11). *Classification interface for WEBSPAM-UK2007.* http://www-connex.lip6.fr/ xmlmining/ws/assesment/cgi-bin/host.php.

WEBSPAM-UK2007, Team. 2007c (11). *Datasets for Research on Web Spam Detection.* http://www.yr-bcn.es/webspam/datasets/.

WEBSPAM-UK2007, Team. 2007d (11). *Guidelines for WEBSPAM-UK2007.* http://www.yr-bcn.es/webspam/datasets/uk2007/guidelines/.

WEBSPAM-UK2007, Team. 2007e. *Web spam challenge homepage.* http://webspam.lip6.fr/wiki/pmwiki.php?n=Main.HomePage.

Weideman, Melius. 2007. Use of Ethical SEO Methodologies to Achieve Top Rankings in Top Search Engines. Cape Peninsula University of Technology, Cape Town, South Africa.

Westbrook, A., & Greene, R. 2002 (Dec). *Using Semantic Analysis to Classify Search Engine Spam.* Class Project report at http://www.stanford.edu/class/cs276a/projects/reports/rdg12-afw.pdf. Stanford University.

Wu, Baoning. 2007. *Finding and fighting search engine spam.* Ph.D. thesis, Bethlehem, PA, USA.

Wu, Baoning, & Chellapilla, Kumar. 2007. Extracting link spam using biased random walks from spam seed sets. *Pages 37–44 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM.

Wu, Baoning, & Davison, Brian D. 2005a (May, 10). *Cloaking and Redirection: A Preliminary Study.* First International Workshop on Adversarial Information Retrieval (AIRWeb).

Wu, Baoning, & Davison, Brian D. 2005b. Identifying link farm spam pages. *Pages 820–829 of: WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web.* New York, NY, USA: ACM.

Wu, Baoning, & Davison, Brian D. 2006a. Detecting semantic cloaking on the web. *Pages 819–828 of: WWW '06: Proceedings of the 15th international conference on World Wide Web.* New York, NY, USA: ACM.

Wu, Baoning, & Davison, Brian D. 2006b. Undue influence: eliminating the impact of link plagiarism on web search rankings. *Pages 1099–1104 of: SAC '06: Proceedings of the 2006 ACM symposium on Applied computing.* New York, NY, USA: ACM.

Wu, Baoning, Goel, V., & Davison., Brian D. 2006 (May). Propagating trust and distrust to demote web spam. *In: Proceedings of the WWW2006 Workshop on Models of Trust for the Web (MTW), Edinburgh, Scotland.*

Xavier-Parreira, Josiane, Castillo, Carlos, Donato, Debora, Michel, Sebastian, & Weikum, Gerhard. 2008. The JXP Method for Robust PageRank Approximation in a Peer-to-Peer Web Search Network. *VLDB Journal.*

Yi-Min Wang, Doug Beck, Xuxian Jiang Roussi Roussev Chad Verbowski Shuo Chen, & King, Sam. Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities.

Zhang, Hui, Goel, Ashish, Govindan, Ramesh, Mason, Kahn, & Van Roy, Benjamin. 2004. Making Eigenvector-Based Reputation Systems Robust to Collusion. *Pages 92–104 of: Proceedings of the*

third Workshop on Web Graphs (WAW). Lecture Notes in Computer Science, vol. 3243. Rome, Italy: Springer.

Zhang, Lei, Zhang, Yi, Zhang, Yan, & Li, Xiaoming. 2006. Exploring both Content and Link Quality for Anti-Spamming. *In: CIT '06: Proceedings of the Sixth IEEE International Conference on Computer and Information Technology*. Washington, DC, USA: IEEE Computer Society.

Zhou, Bin, Pei, Jian, & Tang, ZhaoHui. 2008. A Spamicity Approach to Web Spam Detection. *Pages 277–288 of: SDM*. SIAM.

Zhou, Dengyong, Burges, Christopher J. C., & Tao, Tao. 2007. Transductive link spam detection. *Pages 21–28 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. New York, NY, USA: ACM.

ARMANDO CARVALHO

Este estudo centra-se na imensa informação disponibilizada na Web, procurando melhorias na forma de, com a máxima certeza, ser possível aceder a uma parte dessa imensidão de modo rápido e limpo de interferências estranhas aos conteúdos.

Os motores de pesquisa tornaram-se a "porta principal" para aceder a essa informação, por isso, os seus proprietários desenvolvem cada vez mais esforços preparando algoritmos de alta performance para ajudar os utilizadores a encontrar as páginas com conteúdos mais relevantes para as suas necessidades.

Atraídos pelo crescente número de acessos por estes meios, os *spammers* conseguiram iludir as seguranças criando diversas formas de introduzir informação não desejada no resultado fornecido pelos motores de pesquisa.

Importa, por isso, validar e actualizar os algoritmos de pesquisa, por forma a que resistam a esses engodos, mantendo-se atentos à propria evolução humana para conceitos mais abrangentes.

Os utilizadores têm vindo a aumentar a sua confiança nos motores de pesquisa como um meio de obter informação. Por outro lado, os spammers têm, com êxito, conseguido minar essa confiança adulterando o resultado desejado em cada consulta.

VNiVERSiDAD D SALAMANCA