Cartografias virtuais – mapear o ciberespaço

Sobre os problemas da Recolha de Informação da perspectiva do utilizador[1]

Helena Barbas
CENTRIA e DEP/FCSH Universidade Nova de Lisboa,
Av. de Berna, 26-C, 1069-061 Lisboa, Portugal,
+3517933519 – hebarbas@fcsh.unl.pt

Abstract: Análise da arquitectura da *web*, e do funcionamento dos motores de pesquisa, relativamente às necessidades do utilizador comum. Necessidade de novas teorias e metodologias, novos paradigmas, para avaliar o comportamento humano em contextos de Recolha de Informação.
Analysis of web architecture and browsers performance, in relation with the common user needs. There is a need of new paradigms and perspectives to evaluate human behaviour in IR contexts.

Keywords: Recolha de informação; interacção homem máquina; estudos do usuário;
Information retrieval; human-computer interaction; user studies;

INTRODUÇÃO

O primeiro requisito em todas as propostas de cômputo de sistemas de Recolha de Informação (IR-Information Retrieval) – seja da perspectiva do sistema, seja da perspectiva do utilizador – é a relevância dos documentos numa colecção, e o modo de funcionamento dos motores de pesquisa.

E, caso a avaliação não se reporte a uma base de dados específica (que tem a vantagem de permitir controlar toda a sua arquitectura de raiz [1]), a primeira grande colecção a ser pesquisada é a Internet.

1 ARQUITECTURAS

O princípio que presidiu à invenção da Internet – a possibilidade de criar um espaço de informação não controlado nem controlável – está a tornar-se o seu anátema, por desmesura. Diariamente são-lhe acrescentadas cerca de 15 milhões de páginas das quais, um utilizador normal e assíduo, poderá ler umas 100. Estão em curso várias tentativas para controlar a actual selva do excesso de informação, para que seja de facto útil.

Da parte da própria Web, via W3C, oferecem-se sucessivos documentos que procuram estabelecer uma série de requisitos, limitações e princípios que permitam organizar a «Arquitectura da Web» [http://www.w3.org/TR/webarch/], a que nem sempre se obedece.

[1] Trabalho efectuado no âmbito do Mestrado em Inteligência Artificial, Faculdade de Ciências e Tecnologia, U.N.L.,coord. Pelo Prof. Luís Moniz Pereira – Seminário de Processamento de Língua Natural I – orientado pelo Prof. Paulo Quaresma, 2003. O texto foi revisto em Novembro de 2004, os «links» e referências bibliográficas foram actualizados em 11 de Junho de 2006.

Da parte das instituições estatais e internacionais, com a ONU à cabeça, desenvolvem-se diligências para estabelecer uma norma – ISSS - Information Society Standardization System [2] – que reúna, num único código, as várias propostas já aventadas, mas a quantidade e diversidade destas deixa prever que não seja tão cedo que se consiga um consenso mundial.

Da parte das universidades, pode dar-se o exemplo do trabalho de Parker Rossman [3], que esteve em Lisboa em 1999 a apresentar uma comunicação – «Projectos e mapas: arquitectura para a universidade no ciberespaço» – [4], onde defende um «consórcio» de e-meta-universidades a contribuir para um saber comum e geral, para uma mega-pesquisa – à imagem e semelhança de «The Human Genome Project» [http://www.ornl.gov/hgmis], ou «The Global Knowledge Partnership» [http://www.globalknowledge.org/] – que fundamentasse a criação de uma base de dados global, devidamente indexada.

Há ainda a tentativa de cartografar esse conhecimento e desenhar uma topologia a partir do entendimento da net como uma rede de «routers» associados por «links», em que cada «router» pertença a uma autoridade administrativa ou sistema autónomo (AS). Soon-Hyung Yook, Hawoong Jeong, e Albert-László Barabási [5], descobrem-lhe a topologia física, desenhável por geradores e (relativamente) previsível:

> Ou seja, por mais pormenorizado que seja um modelo da Internet, se os seus parâmetros universais (a, s, D*f)* se desviarem dos que não são cobertos pelas medições, a topologia de larga escala irá inevitavelmente diferir da Internet actual. [6]

Porém a topologia física não coincide, necessariamente, com a quantidade e tipo de informações nela contida.

São muitas as provas da preocupação com o mapear do conhecimento inserido na Web, [veja-se http://www.cybergeography.org/atlas/atlas.html por exemplo] mas a velocidade a que a informação é acrescentada, e a sua quantidade, tornam praticamente impossível estabelecer um quadro fidedigno ou minimamente actualizado. Há até quem brinque e ofereça a possibilidade de se chegar ao «fim» da Internet [http://www.shibumi.org/EotI].

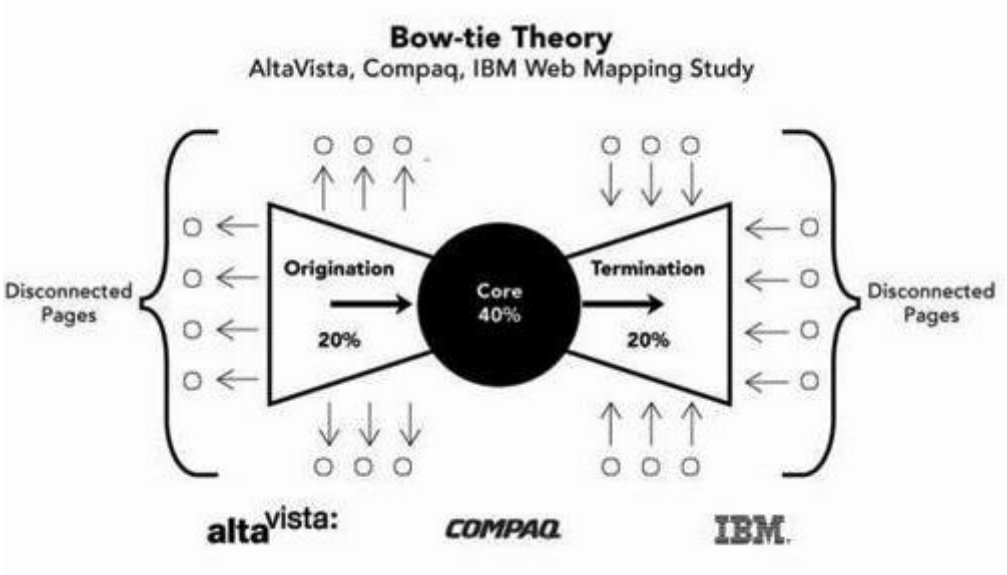
1.1 A rede escondida

Entretanto, as investigações não param. Em Maio de 2002, Chris Sherman divulga um estudo que altera a ideia geral de que a *net* seja constituída por uma esfera de «clusters» de sites bem interligados entre si:

> Um novo mapa do ciberespaço mostra que a Web se assemelha a um laço de pescoço [*papillon*], com limites divisórios que podem tornar difícil ou mesmo impossível a navegação entre regiões, de acordo com um novo estudo publicado por investigadores de AltaVista, Compac e IBM. As teorias prévias sugeriam que a Web estava altamente conectada, com não mais do que 19 graus de separação

> de um «site». Em contrapartida, o novo mapa revela uma estrutura mais subtil que pode levar ao desenvolvimento de técnicas de pesquisa pelos motores (*searchers* e *crowlers*), e uma maior compreensão da sociologia da criação de conteúdos, e isso pode ajudar a prever a emergência de novos fenómenos na Web, como os «Web rings» e «Spam clusters». [7]

Segundo esta teoria, a *net* terá a seguinte a forma:



- «Core» – [Cerne] será o «coração» da Web. As páginas no seu interior estão fortemente conectadas por «cross-linking» [interligações cruzadas]. São os «links» nas páginas do «core» que permitem aos utilizadores viajar com relativa facilidade de umas para as outras; são também eles os mais provavelmente seguidos pelos «browsers»;
- «Origination» – [De origem] páginas que eventualmente permitem ao utilizador atingir o «core» mas que, por si, não podem ser por ele acedidas. Serão páginas obscuras, que ainda não atraíram a atenção da comunidade, ou não têm «links» para as páginas do «core». (ex. Geocities e Tripod);
- «Termination» – [Terminais] páginas a que se pode aceder a partir de «links» no «core» mas que não têm ligação de retorno, são «destinos» em si;
- «Disconected pages» – [Desligadas] podem ser ligadas a páginas «origination»/ «termination», mas não é possível aceder-lhes directamente a partir do «core».

Em termos percentuais, para além do facto de o «core» ser muito menor do que se imaginava, Sherman salienta as seguintes «descobertas»:

1. Para qualquer página de origem ou destino escolhida ao acaso, a probabilidade de que exista um «hiperlink» a partir do «core» é apenas de 24%;

2. Se não existe um «link» directo entre as páginas escolhidas ao acaso, a distância média entre elas é de 16 «links» – ou seja, um «browser» terá que clicar os «links» em 16 páginas para lá chegar. Esta distância é menor que os 19 «links» postulados por estudos anteriores, mas exclui 76% das páginas que não possuem «links» directos.
3. Se existe um «link» indirecto – dos que podem ser seguidos para trás ou para a frente, acessíveis aos «spiders/crawlers», mas não ao utilizador de um «browser» – a média da distância é de 6 «links»;
4. Mais de 90% das páginas da Web podem ser alcançadas umas a partir das outras desde que possuam «links» reversíveis.

Estas conclusões tornam-se importantes na medida em que podem orientar os modos de pesquisa individuais, e mais ainda porque confirmam a existência de grandes zonas de informação a que se não acede com facilidade. Sobre esta «hidden Web» dizia Marcia Mardiz em 2001:

> Actualmente, a WWW contém uma quantidade aproximada de 7.4 milhões de sites (OCLC, 2001). Porém, mesmo o pesquisador mais experimentado, usando o motor de pesquisa mais robusto, só consegue aceder a cerca de 16% dessas páginas (Dahn, 2001). Os outros 84% da informação disponível ao público são chamados de Web «oculta», «invisível», ou «profunda» [8]

Entende, então, que a «web oculta» é 500 vezes maior que a zona indexada pelos «browsers», e nela se encontram documentos, bases de dados e servidores que não usam o html. São constituídas principalmente por cerca de 550 milhões de documentos individuais, pelas informações mais recentes, e conteúdos mais específicos: 95% da informação não sujeita a pagamento ou inscrição prévia. Em termos de conteúdos, serão artigos, imagens de museus, bases de dados, relatórios e trabalhos de investigação especializados (ERIC) – sendo portanto o corpo de informação com maior qualidade.

Não saber ao certo onde, e como, está organizado o conhecimento é a primeira dificuldade a ter em conta quando de uma tentativa de Recolha de Informação. A segunda terá ver com a existência de múltiplos modos de indexação usados pelas diversas colecções.

1.2 Indexação e bases de dados

O denominador comum da comunicação inter-humana é a linguagem natural – escrita e falada. (Cada povo tem a sua – o que pode indicar que o inglês se esteja a tornar o «esperanto» virtual). Todas as outras formas de indexação são auxiliares limitados que os humanos têm primeiro que aprender antes de as poderem utilizar [9].

A maior parte da informação na Web até há pouco era textual. Começa agora a ser superada por som, imagens e filmes. As novas indexações terão ainda que passar a incluir o

processamento de informação digital, electrónica, com linguagens específicas [10]. Isto interferirá naturalmente com os modos como se passa a fazer a pesquisa, levando a que se tenha que aprender como se podem questionar as bases de dados não-textuais [11]. Segundo o Grupo MIRA, a interactividade afecta, naturalmente, a avaliação dos processos de recolha de informação:

> O modelo clássico do sistema de avaliação da Recolha de Informação, iniciado pelas experiências de Cranfield e actualmente manifestas no programa TREC, demonstra muito claramente as suas origens na era de sistemas de recolha por «batch». O sistema é visto como aceitando uma pergunta bem definida («query» ou tópico) e produzindo um resultado bem definido (uma lista de documentos). Porém, com os modernos sistemas interactivos, esse modelo «input-output» está claramente a tornar-se cada vez mais inadequado como representante da situação de RI. Um problema dominante na investigação sobre RI é a questão de saber qual o modelo, ou modelos, de que precisamos para o substituir. Uma possível fonte de ideias e métodos serão trabalhos noutras áreas (fora da RI) para avaliação das características HCI dos sistemas. Porém, estes trabalhos sofrem duas limitações, pelo menos, no que respeita a sua aplicabilidade à RI. [12]

Por sua vez, aquela RI depende da forma como são organizadas as bases de dados. Em princípio, embora na sua categorização e avaliação estejam dependentes do elemento humano [13], são todas HDLs (Hypermedia Digital Libraries [Bibliotecas Digitais Hipermédia]) porque se fundam no paradigma do hipermédia. Existem porém diferenças entre dois tipos básicos, dois sistemas representando, cada, um modelo de interacção, um ambiente de busca de informação, distintos.

No modelo da biblioteca digital hipermédia (WWW) – os utilizadores recorrem a um único interface («browser» [motor de pesquisa]) para aplicar duas estratégias de busca de informação (ISS – Information Seeking Strategies [Estratégias de busca de informação]), «browsing» e «queery» [pergunta] [14]. Embora de utilização mais fácil, nele não existe separação entre os «links» e os documentos por eles referidos; suportam apenas uma estrutura gráfica básica, dando a ilusão de possuir outras como a hierarquização de «clusters» [agrupamentos/índices temáticos] de documentos. Através do «clustering» [15] oferecido pelos «browsers» (ex. Yahoo) [16], podem multiplicar-se as estratégias de RI, mas sempre por intermédio de um mesmo *interface*. Há ainda a hipótese de se recorrer a vários «browsers» em simultâneo e, embora a Web não ofereça nenhum protocolo para interacção e coordenação entre «browsers» [17], existem os meta-pesquisadores [18] que desempenham essa função (o antigo http://www.highway61.com, ou o mais recente http://www.metacrawler.com/, p.ex.). Por seu lado, alguns «browsers» estarão a tentar ultrapassar esta dificuldade, como o Hotbot.Lycos – que se «personalizou» em Janeiro de 2003, passando não só a patentear «filtros», como ainda a permitir a transferência da pesquisa para outros motores: Inktomi, Fast, Google ou Teoma. O segundo modelo (OHS – Open Hypermedia System [Sistema hipermédia aberto]) é uma biblioteca idêntica em termos de

organização e conteúdos, mas que se baseia em Agentes Hipermédia (HA-Hypermedia Agents). Os «links» são armazenados de modo externo e independente dos referentes; são permitidos modelos de informação mais avançados e hierarquizações mais elaboradas. Este sistema de dados proporciona uma forma de interactividade em que se pode recorrer, em paralelo, a múltiplos *interfaces* e diversas estratégias de busca [19]:

1. «browsing» de documentos simples, ou cruzados;
2. «browsing» dos «clusters» – que exibem objectos de informação em bruto;
3. «browsing» hierárquico – mostrando hierarquias de outros compósitos e «clusters»;
4. Pesquisa em índices de conteúdos, suportadas por Agentes Hipermédia de Bibliotecas (HLA – Hipermedia Library Agents), que oferecem uma visão de conjunto do espaço de informação a ser investigado;
5. Busca por «query» em colecções simples e múltiplas, apoiada por um agente especializado de nome *Information Retrieval* (IR) HA.

A diversidade da construção destas colecções implicará, naturalmente, resultados diferentes em termos de pesquisa, e logo, em termos de avaliação de RI. Em sistemas do modelo como OHS as medidas de «Recall» e «Precisão», por exemplo, perdem parte da sua importância [20].

1.3 Tipos de motores de pesquisa

Encontra-se uma legião de estudos [21] sobre os motores de pesquisa, a sua eficácia [22], características e modos de funcionamento [23]. Mas em termos imediatos, apresentam-se como um sistema de recolha de informação que confronta uma pergunta («query») com um índice por si criado (as palavras em cada documento, os indicadores para a respectiva localização dentro dos documentos). Compreende quatro módulos essenciais: um processador de documentos, um processador de perguntas («query»), uma função de busca e comparação, a possibilidade de ordenar hierarquicamente as perguntas/documentos [24]. As avaliações são feitas em termos do âmbito da pesquisa (acesso a dados no maior número possível de zonas [25]), quantidade de documentos auto-indexados, velocidade da recolha [26], e hipóteses de cingir as respostas ao tema da pergunta [27].

De um modo geral, todos os «browsers» estão a funcionar em cima do grande arquivo inicial, que não foi normalizado, sobrepondo-se por vezes em zonas comuns [28]. Todos os dias aparecem novos servidores e portais, desaparecem outros [29], ou fundem-se entre si [30].

Relativamente aos portais, mudou a qualidade dos motores e directórios (Yahoo!) com que se providenciam. O AltaVista perdeu terreno relativamente ao Yahoo, e o Google [31] ultrapassou todos de repente. As capacidades de indexação [32], e o tamanho dos índices a

aumentarem (mas não de forma tão dramática como apregoam [33]) levou a que a inclusão de um «Url» [Uniform Resource Identifier -Identificador uniforme de fonte] passasse a ser paga – uma forma de combater o «spamming» [34] dizem, ou de controlar a adição de urls que em muitos casos era feita manualmente.

Da parte dos «webmasters», refere-se uma maior preocupação em optimizar os «sites», que passa por um maior cuidado no adicionamento dos «urls», nos títulos dados às páginas e descritores usados.

Desde a *Workshop* de Dublim em Março de 1995 [35] que se procura encontrar os tipos de meta-informação correctos e necessários para poder identificar um documento na Web. Estabeleceu-se uma série de elementos descritivos – retirados da catalogação das bibliotecas – considerados essenciais. Mas as categorias usadas para identificar a informação não satisfazem as necessidades mais comerciais da Web, que vão inspirar o «clustering» dos Web browsers – AltaVista, Yahoo e Netscape escolhem «categorias», não coincidentes entre si; Lycos adopta um sistema de «guias» – que além do mais variam de país para país:

> Pode ver-se, a partir desta vasta variedade de esquemas de classificação, que desenvolver motores de pesquisa com base temática que trabalhem através de um âmbito alargado de «sites» na Web, da maneira que os meta-pesquisadores o fazem para a pesquisa de texto livre, não é fazível neste momento. [36]

E acrescentam, no que respeita ao vocabulário e às listas de acrónimos necessários à pesquisa automática:

> Outra área onde é necessária a melhoria, é na preparação de listas de vocabulários e acrónimos para uso com os motores de pesquisa automatizados. Nesta área, o mundo dos padrões da Tecnologia da Informação, veloz e livre de acrónimos, é particularmente ilustrativo. Tentar categorizar os ficheiros de Difuse, que listam as últimas regras para TI, é quase um pesadelo porque nenhuma quantidade de referenciação das listas de vocabulários ou acrónimos existentes pode identificar termos padronizados para referenciar algo que só muito recentemente foi desenvolvido. [37]

Por outro lado, e segundo o estudo de Maria Leonilde Varela [38], o «clustering» de documentos obedece aos mais variados sistemas e, se garante que a informação possa ser obtida em «quantidade», não garante que a essa quantidade corresponda a equivalente qualidade. O mesmo problema se coloca relativamente ao modo de construção dos «tops» por parte dos «browsers». A RI é naturalmente facilitada quando se oferece a possibilidade de escolha de páginas semelhantes (antigo Lycos, e no Google, p. ex.), bem como se quando existe possibilidade de tradução.

Apesar de tudo, os servidores passam a ter maior cuidado com a precisão dos resultados oferecidos, e com o modo como são adicionados os documentos [39].

1.4 Url(s)

A arquitectura da Web começa com uma sintaxe uniforme para identificadores de fontes, a fim de que estas possam ser reconhecidas, para que se tenha acesso a elas, se possa descrevê-las e partilhá-las. A sintaxe URI (Uniform Resource Identifier [Identificador uniforme de fontes]) emprega uma série de esquemas que incorporam protocolos de identificação específicos – «http», «Ftp», «idap», «urn», «tel», «mailto» [40]. Por vezes, as representações são acompanhadas por meta-informação na mensagem (os títulos «http»), o que é fundamental para uma correcta interpretação da fonte, e acaba por condicionar o manuseamento dos identificadores subsequentes.

A um «Url» exige-se, primeiro que, tudo consistência – que todas as fontes sejam devidamente identificadas (que não se use o mesmo «Url» para diferentes fontes, nem que a mesma fonte seja identificada com diversos «Urls»); que a sintaxe esteja correcta e não seja ambígua (não refira um «media» diferente do anunciado) [41]. E persistência: que, no tempo, o mesmo o mesmo endereço refira o mesmo documento [42].

A comercialização da Web leva a que em muitos casos se jogue com esta informação – desvirtuando-a – para reconduzir o utilizador à página de um qualquer produto que nada tem a ver com o procurado (p. ex., o endereço do índice de bolsa Nasdaq, se escrito com K encaminha para um site pornográfico).

O aumento da quantidade de documentos pode ser falseado por dificuldades na gestão dos URI. Há muitas páginas na Web que estão desactualizadas, ou mudaram de «Url», outras são tão recentes que ainda não constam dos índices:

> Tenham pena dos pobres motores de pesquisa. Eles rastejam por este Maëlstrom borbulhante a que chamamos Web, indexando o texto de centenas de milhões de páginas, as quais podem todas mudar de um momento para o outro. Durante os últimos anos, a maioria dos motores de pesquisa afirmava que actualizavam a totalidade das suas bases de dados uma vez por mês, ou à volta disso. Porém, registos mais antigos nas suas bases de dados mostravam que o prazo de actualização era maior do que o afirmado. [43]

Quanto maior for o intervalo entre a visita do «crowler», tanto mais desactualizada estará a informação fornecida, e maior será o número de «links» partidos [44]. A média de actualização dos «browsers», que não é feita em datas certas, pode ir de 1 dia até cerca de quatro meses [45], mesmo no Google [46]. Por sua vez, a percentagem de «links» partidos (em Fevereiro de 2000) variava entre os 14% a 1% [47]. Uma outra experiência (Mike Thelwall) contradiz, em parte, a funcionalidade dos «links» acima aventada por Chis Shermann:

> A investigação também mostra que outros motores de pesquisa importantes podem não responder ao aparecimento de novas páginas na web, mesmo quando estas possuem «links» criados por páginas conhecidas. O resultado não prova

> que a única maneira para que os «sites» seja indexados neste contexto seja o registo directo do «Url»s nos motores de pesquisa, porque é possível que algum aspecto do desenho do «site» de teste o tenha feito ser rejeitado como fonte de novos «Url»s. Também se pode dar o caso de que exista uma acumulação de «Url»s para serem adicionados, tornando o tempo, entre a descoberta de um novo endereço e o necessário espaço no disco para o indexar, superior a sete meses. A secrecidade dos algoritmos utilizados para determinar novos «Url» cria esta incerteza. No entanto, os resultados oferecem um incentivo para registar sítios «web» nos maiores motores de pesquisa, mesmo que os «sites» tenham bons «links» para si. [48]

E acrescenta:

> Quanto aos que se envolvem na recolha de informação por motivos comerciais, académicos ou outros, as diferenças aparentes entre motores de pesquisa é um lembrete de que o uso de um único deles não dá acesso à totalidade da Web. No caso de a informação estar num novo «site», para o qual não existam bons «links», então o pesquisador fica à mercê do conhecimento do «designer» do «web site», ou da sua decisão quanto a registar a página em motores de pesquisa onde a informação nem sequer será encontrada. [48]

Um outro problema que interfere com a RI é o excesso de publicidade – «banners» que demoram a descarregar, janelas «pop-up» que se sobrepõem à informação procurada [49] – e, pior ainda, a publicidade disfarçada que levou a uma lei federal nos U.S.A. sobre a exibição enganosa de «links» pagos [50].

Qualquer página é apresentada a partir do endereço que é o «url», a informação mais visível, mas a ela se acrescentam o descritor, a data, e por vezes a percentagem de interesse relativamente à pergunta feita.

1.5 Tipos(s) de pergunta(s) [Query]

As possibilidades de RI estão dependentes do modo como a pergunta é feita – o que nem sempre depende do utilizador –, e das hipóteses permitidas pelo sistema de «query» oferecido por cada «browser».

É importante a facilidade de uso do *interface*, a sua capacidade para analisar e «compreender» o vocabulário de uma colecção. Por sua vez, o utilizador tem que poder interpretar, e descortinar a relevância da informação fornecida para as suas necessidades do momento. Estes aspectos não são tratados pelos métodos tradicionais de avaliação de RI – que apenas têm em conta a pergunta em si, e os resultados obtidos. Nos casos das bases de dados OHS, a «query» pode servir como ponto de partida para explorar o espaço em busca de informação útil [51].

Na sua maioria, os «browser» oferecem a pesquisa [52] com base na lógica boleana (AND/+, OR, NOT/-, nalguns casos NEAR, parêntesis/nesting, e «Truncation»/*, %), com possibilidade de refinamento ou pesquisa por palavra-chave.

A pesquisa por palavra-chave – embora com bons resultados no que respeita à resposta dos «browsers» [53] – tem sido considerada limitada por se resumir à densidade de palavras, e devolução de um excesso de resultados irrelevantes. O problema está a tentar ser resolvido por duas vias – um sistema em que se desenvolve um léxico do utilizador, que permite determinar o sentido atribuído a uma certa palavra; outro em que o texto de um documento é pesquisado e são analisadas as relações das palavras entre si, a fim de que sejam colocadas em categorias específicas que descrevam melhor a respectiva funcionalidade [54].

Mas estes aspectos não têm em conta o problema principal, que é o facto de os «browsers» alterarem a «query»:

> Os motores de pesquisa muitas vezes interpretam e transformam a pergunta do utilizador durante o processo de recolha. Estes processos afectam profundamente, tanto os resultados da pesquisa, quanto a capacidade do utilizador para compreender as relações entre a pergunta que fez e os resultados que recebeu. [55]

É que cada motor de pesquisa escolhe as suas transformações internas – o que quer dizer que estas variam de uns para os outros – e o utilizador nem sequer se apercebe dessa mudança. Ainda segundo Muramatsu:

> Ao transformar a pergunta do utilizador sem proporcionar qualquer informação quanto a essas modificações, os motores de pesquisa da web interferem com a formação, por parte dos utilizadores, de modelos mentais acurados e contribuem assim para a falta de habilidade do utilizador para encontrar a informação desejada. [56]

E classifica o comportamento dos «browsers» neste ponto, distinguindo entre o tratamento «opaco», «transparente» e «penetrável»:

> O nível «opaco» representa um *interface* que não proporciona qualquer indicação sobre as transformações subjacentes que o sistema executa. «Opaco» é o nível padrão de «feedback» comummente oferecido pelos modernos motores de pesquisa comerciais da web. Em contrapartida, os *interfaces* «transparentes» proporcionam uma indicação visível sobre as transformações que foram aplicadas automaticamente. Por fim, os *interfaces* «penetráveis» proporcionam tanto informação quanto meios para os utilizadores controlarem ou ajustarem as transformações. Assim, concluímos que muitos utilizadores não serão capazes de compreender porque é que, muitas vezes, recebem respostas erráticas e confusas dos motores de pesquisa. Os resultados do estudo sugerem fortemente que a operação «opaca» de transformação das perguntas representa uma barreira substancial para os utilizadores na sua tentativa para compreender como processam as suas perguntas os vários motores de pesquisa. [57]

Mais um aspecto que se torna relevante quando da avaliação em RI pois, até certo ponto, estas «imposições» feitas pelo *interface* ao utilizador podem desviar a pesquisa, aumentar o índice de resultados negativos, e conduzir mais rapidamente à desistência.

Ainda no que respeita aos modos de questionação do «browser», há a ter em conta as possibilidades de serem seleccionados documentos com, ou sem, imagem (na maioria jpg e gif), som (o império do MP3, Wave), etc. [58]. O Altavista oferece uma opção específica, outros permitem escolher apenas as páginas com algum dos elementos acima. Existem ainda «browsers» específicos para o efeito [59].

A tudo o que foi dito, resta acrescentar que as metodologias propostas para avaliação dos sistemas de RI têm sido pensadas em termos «técnicos» – destinadas a informáticos, construtores de «browsers» e bases de dados, só há pouco começando a ter em consideração os problemas do utilizador comum.

1.6 Paradigmas de avaliação

Como se pode perceber, a Recolha de Informação na Web é muito diferente da pesquisa em bases de dados tradicionais [60]. Mas foi na indexação dessas bases de dados (principalmente bibliotecas) que se inspiraram os estudos de RI.

O primeiro paradigma é inaugurado pelos testes de Cranfield I, II e III (Cyril Cleverdon, 1966), em que o investigador faz experiências com colecções de documentos para comparar a eficiência relativa das diversas formas de RI. Lança os conceitos de «Recall» [Recolha] (a fracção dos documentos relevantes recolhidos) e «Precisão» (a fracção de documentos recolhidos que são relevantes) como medidas para os processos de Recolha de Informação. Rijs Bergen (1975) desenvolve o sistema no seu livro *Information Retrieval* (reeditado e revisto em 1979) que se destina essencialmente a estudantes de informática. Como seus herdeiros, já preocupados com o problema do utilizador, embora ainda no espaço das bibliotecas digitais, temos Belkin (1994), Schlichting, C. & E. Nilsen (1996), Saracevic (1995, 1997) e Van House (1995). Neste grupo inclui-se ainda E. Voorhees (2000,2002, 2002) [61]

Em «Reflections on IR Evaluation», Mei-Mei Wu e Diane H. Sonnenwald, fazem uma síntese das teorias e metodologias de RI, a partir de duas perspectivas: do sistema (System-oriented) e do utilizador (User studies).

Referindo-se ao «system-oriented», historiam as diversas tipologias e contribuições, desde a sua origem nos anos 50, com os trabalhos de Cleverdon (UK) – a identificação do algoritmo mais eficiente face a medidas padrão. As pesquisas são alargadas por Salton (USA) para incluir uma avaliação do espaço vectorial em algoritmos de RI.

As investigações vão cristalizar-se em torno da TREC (Text Retrieval Conference – que inspirou a criação de um grupo idêntico no Japão em 1999), no âmbito da qual todos os anos são propostos novos testes, procurando cada um resolver um de 8 problemas diferentes:

- *Cross-Language Track* – descobrir documentos relevantes independentemente da língua;
- *Filtering Track* – o sistema tem que tomar uma decisão binária quanto a novos documentos se são relevantes para os tópicos;
- *Interactive Track* – estudar a interacção do utilizador com os sistemas de RI (relevância de «feedback», p.ex.);
- *Query Track* – efeitos da variação das «queries» e análise da *performance* da recolha;
- *Question answering Track* – estuda a recolha de informação comparada com a recolha de documentos. (O sistema deve responder a 200 perguntas);
- *Spoken document retrieval track* – investiga a capacidade do sistema para recolher documentos orais;
- *Web Track* – 2 GB de informação; se os «links» podem ser usados para melhorar a RI.

As contribuições dos membros da TREC levaram ao refinamento de algoritmos que melhoraram os resultados de «Recall» e «Precisão» em bases de dados de envergadura. Mas suscitam algumas objecções aos autores:

- *validade e fidedignidade*: omitem o usuário, ou relegam-no para um papel passivo; não têm em conta a diferença entre a dinâmica laboratorial e o mundo real; há múltiplos modos de interacção que não são considerados; além de que os julgamentos de relevância não são subjectivos, nem dependentes do tempo-espaço-contexto.
- *possibilidade de generalização*: falta de amostragens relacionadas com usuários; os testes experimentais são demasiado pequenos e limitam-se a tópicos de ciência e tecnologia; as descobertas entram em conflito com o senso comum e a experiência;
- *utilidade*: não são aplicáveis a sistemas operacionais (que têm que ser avaliados por outros critérios); as descobertas são fracas na explicação, predição e controlo dos fenómenos investigados;
- *conceptualização*: não existe um suporte teórico sólido para as medidas e métricas usadas nas abordagens à avaliação de RI; «Recall» e «Precisão» não parecem ter peso significativo para os indivíduos que fazem a pesquisa; as diferenças estatísticas entre estas medidas nos sistemas não são significativas no mundo real, em contexto.

O desenvolvimento das perspectivas que têm em conta o utilizador (User Studies) dá-se pelos anos 70. Têm como pioneiros Saraceviv e Kantor (1988a 1988b), cujo objectivo primeiro é

identificar o comportamento e satisfação do usuário, e equacioná-los com a efectividade do sistema.

As propostas evoluem para incluir pesquisa sobre o comportamento humano – processos e resultados da exploração da informação; busca; filtragem; uso; «provisioning» e disseminação. Centram-se em utilizadores reais (crianças, estudantes, organizações, etc.). Usam métodos e medidas qualitativos (entrevistas, observações, pesquisas) não avaliando explicitamente os sistemas.

Para esta vertente existem grupos análogos ao TREC, a International Conference – Information Seeking in Context, que se transformou num fórum de investigadores que exploram métodos e resultados de pesquisas; a SIG (sobre as necessidades de informação). Diferem entre si porque não promovem uma metodologia de avaliação padronizada.

As suas principais contribuições serão:

- ter identificado o comportamento humano na pesquisa de informação, levando a que se criassem sistemas de RI que incluem interfaces gráficos homem-computador;
- suscitar novos tipos de esclarecimento do usuário sobre as fontes de informação;
- mostrar a necessidade de incluir novos recursos de informação nos sistemas;
- ter ajudado a esclarecer a dinâmica e natureza situacional da «Relevância» – que será contínua e não dicotómica; e que os seus julgamentos são subjectivos, situacionais, e não objectivos ou lógicos (Schamber 1994, determina a «Relevância» como subjectiva, cognitiva, situacional, psicológica, multi-dimensional, dinâmica, e mensurável) [62];

Também relativamente a estes estudos apresentam algumas objecções:

- *não são generalizáveis* - porque, centrando-se numa população específica e pequena, os resultados são contingentes quanto à pessoa, espaço e tempo, logo, não aplicáveis a populações mais vastas ou diferentes;
- *são demorados*;
- *não têm grande utilidade* - não conseguem provocar uma alteração no «design» dos sistemas de RI. Os especialistas de «User Studies» não têm conhecimentos tecnológicos que lhes permitam construir/alterar os sistemas em função dos resultados das suas pesquisas, nem sabem enquadrar ou traduzir os seus resultados de modo a que os construtores de «software» o possam fazer.
- *dificuldades conceptuais* - é difícil sintetizar diferentes níveis e métodos de análise e/ou comparar resultados.

E afirmam: **São necessárias Teorias de comportamento humano na informação que se estendam através de contextos e situações a fim de identificar novas medidas e métodos de avaliação para RI.**

Consideram, pois, ser necessário fazer a ponte entre os entre os diversos paradigmas, e reelaboram as perguntas de Saracevic (1995):

1. que êxito teve/tem a informação recolhida para a solução do problema da explosão de informação nas áreas aplicadas?
2. como pode a RI ajudar as pessoas nas situações em que são confrontadas com problemas de busca (encontrar, usar, e interagir com a informação) frente à massa de informação existente e à miríade de escolhas possíveis?
3. como é que toda essa informação, tecnologia e sistemas de informação afectam o nosso trabalho, os tempos livres, a sociedade e a cultura?

E concluem que estas questões não podem ser respondidas por estudos orientados apenas para o sistema, ou apenas para o usuário. Torna-se necessária uma síntese entre as pesquisas laboratoriais e os contextos, etc., para que se possa demonstrar a eficácia a partir da perspectiva do utilizador.

Discutindo os enquadramentos de avaliação relacionados com a RI, reconhecem que alguns deles vêm de disciplinas da área das ciências humanas e cogntivas: métodos e técnicas para avaliar a proficiência do *interface* homem-computador; a dimensão do trabalho cognitivo e tipo de estratégias mentais; etc.

Sintetizando enquadramentos e paradigmas, vão adaptar os cinco atributos que Rogers (1995) relaciona com a inovação:

- *vantagem relativa* - até que ponto uma inovação substitui as práticas correntes, é tornada operacional, pode ser medida em termos de variáveis (economia, ganho, maior conveniência e prestígio social);
- *compatibilidade* - até que grau a inovação é percebida como consistente com os valores, experiências dos usuários e suas necessidades futuras (por comparação com a estrutura social, crenças individuais e de grupo, clima organizacional ou social, objectivos do indivíduo ou do grupo);
- *complexidade* - dificuldade em aprender a usar e compreender um novo sistema ou tecnologia (medida pelo número de novas destrezas e/ou conhecimentos que é necessário adquirir para usar e beneficiar da inovação);

- *ser testável* - facilidade de experimentação com a inovação numa base limitada (nível de esforço necessário e risco envolvido na observação e participação em demonstrações em pequena escala);
- *ser observável* - grau pelo qual os resultados da inovação são observáveis;

Considerando que alguns atributos podem parecer mais importantes do que outros, propõem que sejam todos usados como fundamento para um novo sistema de avaliação, tendo em conta que nem todas as medidas e critérios são necessários para todas as situações. Oferecem pois, um novo paradigma:

Atributo	Critério	Medidas
vantagem relativa	Relevância do sistema Relevância dos tópicos Velocidade Ganho económico	*Recall* e Precisão Conteúdos da fonte (tipo e cobertura) Tempo de resposta do sistema Análise do benefício de custos
compatibilidade	Relevância motivacional Relevância organizacional Relevância social	Corresponde às expectativas: - do usuário - da organização - sociais - da política pública
complexidade	Usabilidade Relevância cognitiva Relevância situacional	Tempo de completação da tarefa, *ratio* de erros, tempo de correcção de erros, Satisfação do usuário: - em contexto de trabalho - na resolução de problemas
ser testável	Facilidade de experimentação	Disponibilidade, tempo de treino, outros custos de lançamento
ser observável	Grau de demonstração	Custo da observação

Uma variedade de técnicas pode ser usada para calcular as medidas, incluindo experiências laboratoriais de RI para medir «Recall», «Precisão», e tempo de resposta do sistema; outras engenharias para medir tempo de completação das tarefas, *ratio* de erro, tempo de correcção de erro, e tempo de treino; combináveis com entrevistas, observações e pesquisas para determinar a satisfação do utilizador.

Todavia, embora se preocupe com o contexto e a situação em tempo real, mesmo este novo paradigma deixa de fora as estratégias de pesquisa a que qualquer utilizador recorre, que pertencem ao campo da «Relevância» [63], mas são condicionadas pelo objecto que tem diante de si, pelas possibilidades que lhe oferece o interface.

3 DA PERSPECTIVA DO UTILIZADOR

Num dos poucos artigos dedicados à análise das funções de RI da perspectiva do utilizador, Thorsten Joachims (2002) [64] propõe vários testes para determinar a «Retrieval performance user centered» [Sucesso de recolha centrada no usuário]. Mas os seus raciocínios enfermam de

algumas deficiências porque, na construção do quadro-teste 1, mistura «rankings» de browsers diferentes (Google e MSNSearch) sem ter em conta o modo como cada um hierarquiza as respectivas listagens; e na análise dos «links» escolhidos pelo utilizador, nesse quadro, não considera a importância do nem do «url», nem do descritor – que denunciam o interesse do usuário-cobaia por «software».

Há também uma série de manuais [65] e artigos que se preocupam com o estudo dos comportamentos [66] e procuram ensinar truques de pesquisa aos utilizadores, a maioria centrando-se nos modos de formular as perguntas e aproveitar as possibilidades de refinamento de busca. E alguns preocupam-se com as informações fornecidas pelos títulos [67]. Sobre os «urls», há um artigo que aborda o modo como os «crawlers» os visitam e recolhem [68], e um outro sobre o modo como podem ser interpretados [69]. Mas em nenhum dos textos consultados foi tida em conta a possibilidade de o «url» influenciar a decisão do utilizador.

Mesmo sem grandes explicações, com o hábito, qualquer utilizador acaba por olhar para o «url» como factor de decisão no momento de abrir um qualquer documento, dado que acaba por perceber quais os descritores que lhe podem oferecer alguma qualidade.

Assim, no formato:

http://www.server.domínio/directório/sub-directório/nome-de-ficheiro/tipo-de-ficheiro

para além de destrinçar os usuais «org», «com» ou «gov» que já entraram no conhecimento comum, caso o domínio se apresente com «edu», o utilizador saberá que o documento pertence a uma universidade, em princípio uma instituição fidedigna. E entre um documento com um qualquer nome, ou aquele descritor, naturalmente escolherá esse. O mesmo se coloca relativamente a directórios e sub-directórios. Quanto ao tipo de ficheiro, também é fácil perceber se se trata de um documento em html, pdf, ou imagem.

Uma outra estratégia, quando há interesse por um documento que apresenta o «link» partido, e o «browser» não lhe indica o «cache», é ir apagando os sucessivos apêndices até chegar à página inicial. Mesmo sem ser iniciado em informática, o utilizador acaba por construir uma série de estratégias próprias para navegar pelo turbilhão de informações sem mapa com que é confrontado.

H.B., Novembro 2004

Anexo:

Tendo em conta as estratégias de «sobrevivência» na navegação pela *net*, e relativamente aos modos de RI por parte do utilizador, poderia elaborar-se uma lista de perguntas que permitissem construir um inquérito para determinar estatisticamente comportamentos-tipo, por parte do utilizador. Por exemplo:

- Se recorre à *net* em primeira instância (busca inicial às cegas/ em bruto) ou última instância (busca orientada) [70]
- Qual o «browser» preferido e porquê
- modo como faz a pergunta...
    - utiliza estratégias de refinamento
    - recorre a alguma outra estratégia particular
- modo como recebe documentos apresentados
    - se são muitos
    - se são poucos
    - faz pesquisa interna
    - consulta todos indiscriminada ou sucessivamente
    - tem em consideração os «tops» oferecidos pelos browsers
- como lida com a hierarquização das respostas apresentadas
- qual a média de consultas até alcançar um resultado satisfatório
- perde-se/deambula
- a partir de que quantidade desiste de consultar
- o que pode motivar a desistir mais depressa
    - cansaço
    - links partidos
    - frustração nas respostas
    - lentidão de abertura das páginas
    - publicidade
- Modo como lê informação recebida
- qual a primeira opção para abrir um documento
    - título
    - descritor
    - percentagem
    - data
    - endereço
- Hierarquizar a importância de cada uma destas informações
- estratégias para garantir a fidedignidade da informação
- uso e ou citação de todos os documentos
    - só dos «fidedignos»
    - como os reconhece [71]

Notas

[«Links» confirmados e/ou actualizados em 11.Jun.2006]

[1] Paulo Quaresma, *Agentes Inteligentes para sistemas de pesquisa de informação de textos*, D.I.U.E./Centria, 2001, http://www.di.uevora.pt/~pq/brasil/curso/cursoBrasil.htm ;
[2] http://www.cenorm.be/cenorm/businessdomains/businessdomains/isss/index.asp;
[3] Parker Rossman, http://ecolecon.missouri.edu/globalresearch/author.html;
[4] Parker Rossman, colóquio «*Da Ideia de Universidade à Universidade de Lisboa*», http://cie.fc.ul.pt/seminarioscie/universidade/index.html;
[5] Soon-Hyung Yook, Hawoong Jeong, Albert-László Barabási *Modeling the Internet's Large-Scale Topology*, *Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA*, http://www.nd.edu/~alb/cv.html;
[6] Soon-Hyung Yook, Hawoong Jeong, Albert-László Barabási: *«That is, no matter how detailed an Internet model is, if its universal parameters (a, s, Df) deviate from those uncovered by measurements, the large-scale topology will inevitably differ from the current Internet.», Ibid.;*
[7] Chris Sherman: *«A new map of cyberspace shows that the Web resembles a bow tie, with divisive boundaries that can make navigation between regions difficult or even impossible, according to a new study published by researchers at AltaVista, Compaq, and IBM. Previous theories suggested that the Web was highly connected, with no more than 19 degrees of separation from any one site to another. By contrast, the new map reveals a subtler structure that may lead to more efficient search engine crawling techniques and a greater understanding of the sociology of content creation, and that may help predict the emergence of new phenomena on the Web such as Web rings and spam clusters.»* in *The Invisible Web, New Web Map Reveals Previously Unseen 'Bow Tie' Organizational Structure*, in *Information Today*, 22.Maio 2000, in http://www.infotoday.com/newsbreaks/nb000522-1.htm;
[8] Marcia Mardis: *«Currently, the World Wide Web contains an estimated 7.4 million sites (OCLC, 2001). Yet even the most experienced searcher, using the most robust search engines, can access only about 16% of these pages (Dahn, 2001). The other 84% of the publicly available information on the Web is referred to as the "hidden", "invisible," or "deep" Web.»*, in «Uncovering the Hidden Web, Part I: Finding What the Search Engines Don't » Outubro 2001 - EDO-IR-2001-02; http://www.ericdigests.org/2002-2/web.htm;
[9] AIIM International, *Capture, Indexing & Auto-Categorisation Intelligent methods for the acquisition and retrieval of information stored in digital archives* - resumo em: http://www.project-consult.net/Files/Summary%20IWP2103-engl.pdf;
[10] http://www.cenorm.be/cenorm/businessdomains/businessdomains/isss/index.asp;
[11] First Annual Diffuse Conference, From Convergence to Consolidation - *What's Next in the Information Market?* A conference organized by the IST Diffuse Project - 7. Março. 2001, Bruxelas, http://lists.w3.org/Archives/Public/www-annotation/2001JanJun/att-0001/01-Diffuse-Event.html;
[12] Grupo «MIRA»: *«The classical model of IR system evaluation, initiated by the Cranfield experiments and currently manifest in the TREC programme, demonstrates very clearly its origins in the era of batch retrieval systems. The system is seen as taking well-defined input (a query or topic) and producing well-defined output (a list of documents). However, with modern interactive systems, that input-output model is clearly becoming more and more inadequate as a representation of the IR situation. A dominant problem in current IR research is the question of what model or models we need instead. One possible source of ideas and methods is work elsewhere (outside IR) on evaluating the HCI characteristics of systems. However, this work suffers from two limitations, at least as regards its applicability to IR.»* in «Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications», http://www.dcs.gla.ac.uk/mira/themes2.html#Media;
[13] Virtual Resource Centre, http://www.virtuallrc.com/about.html;
[14] Michail Salampasis & Konstantinos I. Diamantaras, **«**Experimental User-Centered Evaluation of an Open Hypermedia System and Web Information Seeking Environments», 2002, http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Salampasis/;
[15] Maria Leonilde Varela, «Clustering de Documentos», Relatório para a disciplina de Processamento de Língua Natural I, (1999-2000), http://www.di.uevora.pt/~pq/miaa/cluster_miaa.html;
[16] Michail Salampasis & Konstantinos I. Diamantaras, *Op. Cit.;*
[17] *Ibid.;*
[18] http://www.virtuallrc.com/about.html;
[19] *Ibid.;*
[20] Grupo MIRA, «Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications», http://www.dcs.gla.ac.uk/mira/themes2.html#Media;
[21] VV.AA. Evaluation of information sources, http://www2.vuw.ac.nz/staff/alastair_smith/evaln/evaln.htm;
[22] Greg R. Notess, «Search Engine Showdown Reviews», Mar. 03, 2003, http://www.searchengineshowdown.com/reviews/;
[23] Greg R. Notess, «Search Engine Features Chart», Jan. 21, 2003, http://www.searchengineshowdown.com/features/;
[24] Elizabeth Liddy, «How a Search Engine Works», Maio, 2001, http://www.infotoday.com/searcher/may01/liddy.htm;
[26] Anne Clyde, **«**Search Engines: An Overview», Teacher Librarian, vol. 27, nº.4, Abril 2000, pp.22-28;

[27] Diana Botluk, «Search Engines Comparison», Law Library Resource Xchange, LLC.2001, http://www.llrx.com/features/engine2001.htm;
[28] Greg Notess, «Search Engines Statistics: Database Overlap», 6.Março, 2002, http://www.searchengineshowdown.com/statistics/0002overlap.shtml;
[29] Greg Notess, «Dead Search Engines»**,** *ONLINE* vol. 26, nº. 3, Maio/ Junho 2002, http://www.onlinemag.net/may02/OnTheNet.htm;
[30] Greg Notess, «Browser diversity», *ONLINE*, Julho 2001, http://www.onlinemag.net/OL2001/net7_01.html;
[31] Greg R. Notess, «Search Engine Statistics: Relative Size Showdown Data from search engine analysis» - 31.Dez. 2002, http://www.searchengineshowdown.com/statistics/9901size.shtml;
[32] Greg R. Notess, «Search Engine Statistics: Database Total Size Estimates» - Dec. 31, 2002, http://www.searchengineshowdown.com/statistics/sizeest.shtml;
[33] Greg R. Notess, «Search Engines Statistics: Database Change Over Time» - Dec. 31, 2002, http://www.searchengineshowdown.com/statistics/change.shtml;
[34] http://websearch.about.com/library/weekly/aa021803a.htm e http://websearch.about.com/library/weekly/bl-seo101.htm?PM=ss11_websearch;
[35] The Diffuse Project - the European Commission's Information Society Technologies programme http://www.hi-europe.info/files/2000/diffuse.htm; Diffuse publications are maintained by TIEKE (the Finnish Information Society Development Centre), http://www.tieke.fi/in_english/about_tieke/ IC Focus http://pi.ijs.si/PiBrain.exe?Cm=Org&Org=IC+FOCUS+LIMITED, The SGML Centre: http://www.infoloom.com/gcaconfs/WEB/ts2055/tp2055.HTM e http://www.hi-europe.info/files/2000/diffuse.htm;
[36] *«It will be seen from this wide variety of classification schemes that developing subject-based search engines that will work across a wide range of web sites, in the way that meta-search engines do for free-text searching, is not currently feasible.», Ibid.*
[37] *«Another area where improvement is required is in the preparation of vocabularies and acronym lists for use with automated searches. In this area the fast moving, acronym-ridden, world of IT standards is particularly illustrative. Trying to categorize the Diffuse files, which list the latest standards for IT, is somewhat of a nightmare as no amount of referencing of existing vocabularies or acronym lists will identify standardized terms for referencing something that has only recently been developed.», Ibid.;*
[38] Maria Leonilde Varela, *Op. Cit.*
[39] Sergey Brin e Lawrence Page, «The Anatomy of a Large-Scale Hypertextual Web Search Engine», http://www-db.stanford.edu/~backrub/google.html ;
[40] «Architecture of the World Wide Web W3C Working Draft 15», Novembro 2002, http://www.w3.org/TR/2002/WD-webarch-20021115/;
[41] *Ibid.*
[42] R. Petke, «Registration Procedures for URL Scheme Names», 1999, http://www.w3.org/Addressing/;
[43] Greg Notess: *«Pity the poor search engines. They crawl this seething, bubbling maelstrom we call the Web, indexing the text from hundreds of millions of pages, all of which can change at a moment's notice. For the past few years, most search engines claimed to refresh their entire database once a month or so. Yet, older records in their databases showed that the refresh rate was often more than claimed rate»,* in *Freshness Issue and Complexities with Web Search Engines»* in, *ONLINE*, Information Today, Inc, 2001, http://www.onlinemag.net/OL2001/net11_01.html;
[44] Greg Notess, «Freshness Issue and Complexities with Web Search Engines», 2001, http://www.onlinemag.net/OL2001/net11_01.html;
[45] Greg Notess, «Search Engine Statistics: Freshness Showdown», 20 Out.2002, http://www.searchengineshowdown.com/statistics/0210freshness.shtml;
[46] http://www.webmasterworld.com/forum3/2657.htm;
[47] Greg Notess, «Search Engine Statistics: Dead Links Report», 21.Fev. 2000, http://www.searchengineshowdown.com/statistics/dead.shtml;
[48] Mike Thelwall: *«The survey also shows that other important search engines can be unresponsive to the appearance of new web pages, even if these are linked to by known pages. The evidence does not prove that the only way to get sites indexed in this context is by registering the URL directly with the search engines because it is possible that some aspect of the test site design caused it to be rejected as a source of new URLs. It may also be the case that there is a backlog of URLs to be added, making the time between finding a new URL and having the free disk space to index it longer than seven months. The secrecy of the algorithms used to determine new URLs creates this uncertainty. The results do, however, provide an incentive to register web sites in major search engines, even if sites are well linked to.» e «For those engaged in information retrieval for commercial, academic or other reasons the apparent differences between search engines is a reminder that the use of a single search engine does not give access to the entire web. In the case where the information was likely to be on a newer site that is not well linked to then the information retriever is at the mercy of the web site designer's knowledge or decision about whether to register their site in search engines as to whether the information is findable at all.»* in «The Responsiveness of Search Engine Indexes» in *Cybermetrics - International Journal of Scientometrics, Informetrics and Bibliometrics,*

ISSN 1137-5019 Vol. 5 (2001) Issue 1. Paper 1. - 22.Fev./2001, http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html ;

[49] *Chip Bayers*, «I'm Feeling Lucky», *Wired*, issue 9.10 – Out. 2001, http://www.wired.com/wired/archive/9.10/google.html ;

[50] Stefanie Olsen, «Search sites work to clean up their act», *CNET News.com*, 19.Agosto, 2002, «*The commercial practices of search engines are once again in the spotlight after a recent warning shot from federal regulators over inadequate disclosure of paid links.*», http://news.com.com/2100-1023-954171.html?tag=fd_top;

[51] «Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications» http://www.dcs.gla.ac.uk/mira/themes2.html#Media;

[52] Diana Botluk, «Search Engines Comparison in Law Library Resource Xchange», *LLC*.2001, http://www.llrx.com/features/engine2001.htm;

[53] Greg R. Notess, «Search Engine Statistics: Unique Hits Report», 6.Març., 2002, http://www.searchengineshowdown.com/statistics/unique.shtml;

[54] Andrew Goodman, «Meaning-Based Search Redefines Web Sleuthing» in *Metaguide* #6.3, 2000, http://www.traffick.com/story.asp?StoryID=57 ;

[55] Jack Muramatsu e Wanda Pratt: *«Search engines often interpret and transform a user's query during the retrieval process. These processes profoundly affect both the search results and the users' ability to understand the relationship between their query and the returned results.»* in **«**Transparent Queries: Investigating Users' Mental Models of Search Engines**»,** *Information & Computer Science*, University of California, Irvine, *SIGIR'0* 1, 9-12 Set. 2001, Nova Orleans, Louisiana, USA. Copyright 2001 ACM 1-58113-331-6/01/0009, http://portal.acm.org/citation.cfm?id=383991&dl=ACM&coll=ACM;

[56] *«By transforming the user's query without providing any feedback on those modifications, web search engines interfere with users' formation of accurate mental models and thus contribute to the users' inability to find the desired information. », Ibid.;*

[57] *«opaque level represents an interface that does not provide any indication of the underlying transformations that the system performs. Opaque is the standard level of feedback commonly provided by modern commercial web search engines. In contrast, transparent interfaces provide visible feedback on the automatically applied transformation. Finally, penetrable interfaces provide both feedback as well as a means for users to control or adjust the transformations. Thus, we conclude that many users will not be able to understand why they often receive erratic and confusing results from search engines. The study results strongly suggest that the opaque operation of query transformations represent a substantial barrier for users in their attempt to understand how various search engines process queries.» Ibid.;*

[58] Phil Bradley, «Finding images on the Internet», 2000, http://www.philb.com/findimages.htm**;**

[59] Mary Colette Wallace, «The Science and Art of Online Research in the Fine Arts: A Process Approach», 2001, http://www.infotoday.com/searcher/sep01/wallace.htm;

[60] Jacek Gwizdka e Mark Chignell, «Towards Information Retrieval Measures for Evaluation of Web Search Engines»,1999, http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf;

[61] Voorhees, E. (2002), «The Philosophy of Information Retrieval Evaluation», To appear in *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum*, CLEF 2001, Darmstadt, Germany, http://www.itl.nist.gov/iad/894.02/works/papers.html; Buckley, C. & Voorhees, E. (2000, Julho), «Evaluating Evaluation Measure Stability» *in Proceedings of the 23rd Annual International ACM SIGIR* Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 33-40, http://www.itl.nist.gov/iad/894.02/works/papers.html; Voorhees, E. (Setembro 2001) «Evaluation by Highly Relevant Documents» in *Proceedings of the 24th Annual International ACM SIGIR* Conference on Research and Development in Information Retrieval, Nova Orleans,LA, USA, pp. 74-82, http://www.itl.nist.gov/iad/894.02/works/papers.html;

[62] São listados 80 factores distribuídos por seis categorias/listas parciais de variáveis que afectam a «Relevância» – julgamento, pedidos, documento, sistemas de informação, condições do julgamento e escolha de escala.

[63] Em Saracevic (1996) – a «Relevância» indica uma relação, que se pode manifestar em cinco tipos: 1. «Relevância» do sistema ou algorítmica – relação entre a «query» e objectos de informação no ficheiro de um sistema quando recolhidos, ou na falta de serem recolhidos, por um dado procedimento ou algoritmo; eficácia comparativa é a relevância inferida é o critério para a relevância do sistema; 2. «Relevância» tópica ou de assunto – refere-se à relação ente o assunto ou tópico expresso numa pergunta, e o tópico ou assunto coberto pelos textos recolhidos, pelos textos no sistema de ficheiros, ou existentes. «Aboutness» – é o critério pelo qual a topicalidade é inferida; 3. «Relevância» ou pertinência cognitiva – refere-se à relação entre o estado de conhecimento e a informação cognitiva necessitada pelo usuário e os textos recolhidos, seja num ficheiro, ou que existem; «Relevância» cognitiva é inferida por critérios de correspondência cognitiva, quantidade e qualidade de informação, novidade, etc. 4. «Relevância» situacional ou utilidade – relação entre a situação, tarefa, ou problema em causa, e os textos recolhidos, os textos no sistema de ficheiros, ou existentes. A «Relevância» é inferida por critérios como utilidade na tomada de decisão, adequação da informação para a resolução de um problema, redução da incerteza; 5.«Relevância» motivacional ou afectiva – refere-se à relação entre as intenções, os objectivos e motivações de um utilizador, e os textos recolhidos, os textos no sistema de ficheiros, ou existentes. Os critérios para avaliação desta relevância são satisfação, êxito, realização. O principal problema é que torna difícil comparar a eficácia de

diferentes sistemas de RI (só usando o mesmo grupo em vários sistemas, ou arranjando outras medidas de avaliação), http://www.scils.rutgers.edu/~tefko/articles.htm;
[64]Thorsten Joachims, «Engines using clickthrough data/ Evaluating Retrieval Performance using clickthrough data», Fev.2002, http://www.cs.cornell.edu/People/tj/publications/joachims_02b.pdf;
[65] Greg Notess, «Learning About Searching», http://www.searchengineshowdown.com/strat/;
[66] «Ethnomethodology and the Evaluation of Information Retrieval Systems: Abstract», Clare F. Harvey, School of Computing and Information Systems, University of Sunderland., 1998, http://www.dcs.gla.ac.uk/mira/workshops/grenoble/harvey.html;
[67] Greg Notess, «Title Searching Showdown», Maio , 2002, http://www.searchengineshowdown.com/features/title/ e «Tracking Title Search Capabilities», *ONLINE*, Maio 2002, http://www.onlinemag.net/OL2001/net5_01.html; David P. Habib e Robert L. Balliot, *How to Search the World Wide Web: A Tutorial for Beginners and Non-Experts*, 1999, 2000, http://amby.com/tools/search.html e US Dep. of Education, http://www.ed.gov/search/searchhelpHow.jsp;
[68] Junghoo Cho, Hector Garcia-Molina, Lawrence Page, «Efficient Crawling Through URL Ordering», 1997, http://oak.cs.ucla.edu/~cho/papers/cho-order.pdf;
[69] Russ Haynal, «How to Read a URL», 1999, http://navigators.com/url.html;
[70] Sim d'Hertefelt, «The Skeptical Internet User Does Not Search», *Interaction Architect*, Nov.2000, http://www.interactionarchitect.com/articles/article20001122.htm;
[71] VV.AA, *An Educators' Guide to Credibility and Web Evaluation*, 1999-2002, http://lrs.ed.uiuc.edu/wp/credibility/index.html;