Instituto Superior Técnico

Bibliotecas Digitais

2005/2006

RECOLHA - IST

48278 – João Miranda

Índice

1. Introdução

Com este trabalho pretende-se efectuar a utilização e teste do sistema RECOLHA da Biblioteca Nacional Digital (BND), através do desenvolvimento de um serviço de arquivo de todos os sítios do IST. O objectivo primário prende-se com a função de teste associada ao espaço **.ist.utl.pt* que permite avaliar as capacidades do sistema e apontar eventuais falhas detectadas e melhorias a efectuar.

1.1. O Sistema RECOLHA

Não é objectivo deste relatório explicar a fundo o funcionamento do sistema RECOLHA, pelo que apenas é dada uma explicação genérica do seu funcionamento. O sistema é activado através de *daemons* que são corridos em linha de comandos e que lançam os processos de recolha necessários. Os *daemons* correm automaticamente em ciclo e vão fazendo a gestão das operações entretanto ordenadas pelo utilizador. A interacção com o sistema faz-se através de uma interface onde é possível gerir as recolhas e onde, por exemplo, se podem criar colecções, adicionar recursos ou definir processos de recolha (Figura 1).

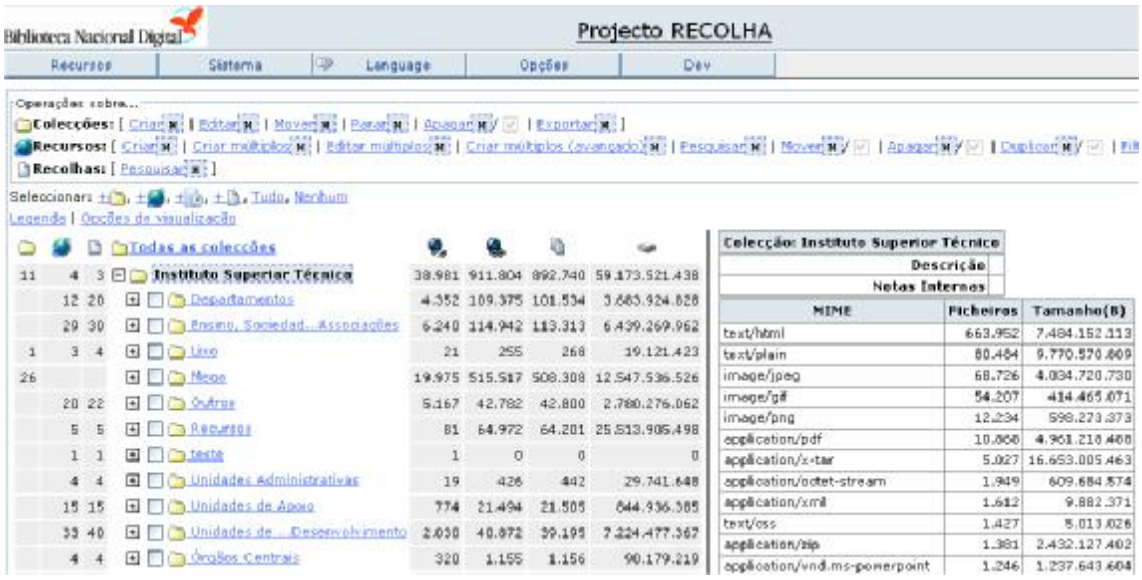

Figura 1- Interface do Sistema RECOLHA

2. Estruturação do Problema

Inicialmente, foi feita uma demonstração do sistema RECOLHA na Biblioteca Nacional com a explicação do sistema pelos responsáveis pelo projecto RECOLHA da BND, mais precisamente pelo engenheiro João Luzio, que viria mais tarde a auxiliar na instalação do sistema em servidor acessível no INESC-ID.

De seguida foram identificados os endereços dos serviços e sítios do IST na *internet*. Como ponto de partida, foi utilizada inicialmente a página principal do IST que contém uma lista razoável de ligações para páginas do Técnico. Esta lista foi complementada com endereços previamente conhecidos e com a pesquisa de endereços nas páginas residentes no domínio do IST e em motores de busca.

Foram identificados cerca de 120 endereços referentes a departamentos, centros de investigação, unidades de apoio e administrativas, e órgãos centrais, entre outros. Foram ainda identificados cerca de 9400 endereços de utilizadores da máquina `mega` em *mega.ist.utl.pt*.

Depois de alguns testes iniciais, as colecções de recursos a recolher foram carregadas no sistema, com a configuração julgada adequada a cada caso. Foi feito o acompanhamento dos processos de recolha, com a correcção de erros entretanto surgidos e melhoramento das recolhas efectuadas.

3. Dificuldades Encontradas

A evolução do sistema é, em geral, lenta e é condicionada por diversos factores como o número de recursos adicionados ou o desempenho da rede a que está ligado o computador onde o sistema está instalado. Em determinadas alturas, registaram-se perturbações de rede que limitaram o processo de recolha dos recursos e gestão do sistema.

Os recursos foram progressivamente adicionados e o desempenho do sistema era afectado à medida que o número de recursos aumentava. O número elevado de recursos que foram adicionados fizeram com que o sistema ficasse sobrecarregado e, por vezes, não respondesse como desejado. A sobrecarga do sistema originava a que determinadas ordens dadas ao sistema só fossem cumpridas dias depois, uma vez que todos os processos pendentes teriam de ser resolvidos entretanto.

O cálculo de estatísticas, já de si com um grande peso nos recursos do sistema, foi condicionado pelo elevado número de recursos existentes. Em determinada fase do trabalho, o processo que efectuava o cálculo de estatísticas originava um erro de falha de memória, que veio a ser corrigido mais tarde com o aumento da memória reservada para o efeito. Este erro fazia com que o sistema ficasse suspenso e depois fosse necessário reiniciar outra vez para que começasse a recolher o que estava em espera. As estatísticas não eram actualizadas e o sistema voltava a ficar suspenso decorridas algumas horas. A solução encontrada foi aumentar a memória reservada para o efeito e correr apenas os processos normais sem o cálculo de estatísticas, uma vez que a recolha efectuada ao mesmo tempo que o cálculo das estatísticas representa uma carga adicional para o sistema. Desta forma, para calcular as estatísticas apenas era corrido o processo de cálculo de estatísticas, sendo desligados os restantes processos. Houve ainda outro erro que não deixava terminar o cálculo das estatísticas até ao fim mas que não foi resolvido em tempo útil.

Fora do âmbito das estatísticas, quando surgia uma excepção no `Httrack` (Figura 2), o sistema ficava por vezes suspenso e era então necessário reiniciar o sistema para recomeçar a recolher o que havia ficado em espera.

Inicialmente constatou-se que certos recursos não recolhiam determinada parte das ligações existentes. Por exemplo, na página do DEI em *http://www.dei.ist.utl.pt* era tudo recolhido menos a parte referente ao pessoal docente. Verificou-se que a ligação referente ao pessoal docente remetia para o endereço *http://www.dei.ist.utl.pt:8080/pessoal/*, pelo que foi necessário ordenar ao sistema que recolhesse todas as ligações que encontrasse a partir do endereço principal (+**www.dei.ist.utl.pt**). Ocorrências deste género verificaram-se noutros recursos, onde determinadas ligações apontavam para servidores específicos dentro do servidor principal e que eram desconhecidos à partida. Foram rectificadas as recolhas relevantes

onde esta situação se verificava e os recursos entretanto adicionados já foram submetidos tendo em conta a recolha de todos os subdomínios referentes a esse recurso.

Figura 2 - Excepção no Httrack

Em relação aos utilizadores do mega, as ligações haviam sido originalmente introduzidas na forma *mega.ist.utl.pt/~user*, sem a barra final, e depois o sistema considerava todo o mega como a mesma directoria e recolhia, por cada um dos recursos, todas as ligações existentes nesse recurso para endereços do mega, resultando em recolhas repetidas e de dimensões indesejadas. Foram redefinidos todos os recursos correspondentes aos utilizadores do mega, com a barra final nos endereços, e as recolhas passaram a ser efectuadas como pretendido.

Não foi possível recolher com sucesso a página do Fénix (*fenix.ist.utl.pt*). Inicialmente estava a ser recolhido o Fénix em conjunto com a página principal do IST (*www.ist.utl.pt*). No entanto, as páginas são na esmagadora maioria recolhidas em branco, com tamanho a zero *bytes*, logo no início do varrimento, pelo que a lista de ligações a recolher acaba prematuramente. O Fénix funciona em `http` e `https` mas redirecciona sempre para `https`. Pensou-se que seria um problema relacionado com o tipo de suporte do `https` por parte do `Httrack`, mas houve outros recursos em `https`, como o TDI (*https://tdi.tagus.ist.utl.pt/*), cuja recolha foi efectuada sem problemas. Pensou-se ainda que pudesse ser um problema relacionado com os *aliases* do Fénix e definiu-se o Fénix como um recurso próprio a recolher a partir de *https://fenix.ist.utl.pt/publico/showDegreeSite.do?method=showDescription.* O resultado das recolhas foi idêntico. Concluiu-se que o problema deve ter origem no próprio Fénix e que deve derivar do mesmo problema que fazia com que os documentos de apoio das cadeiras armazenados no Fénix fossem constantemente descarregados a zero *bytes* e que obrigava os docentes a disponibilizar os ficheiros de apoio das cadeiras no `mega`.

A recolha do sítio do Departamento de Matemática não foi bem sucedida. As inúmeras tentativas de recolha originavam um erro no `Httrack` e a recolha terminava pouco depois de iniciada. Houve uma tentativa, a última, em que o erro não terminou a recolha, mas quando haviam sido recolhidos mais de oito *gigabytes* houve um erro diferente no `Httrack` (Figura 3) e a recolha efectuada, em vez de se manter, foi automaticamente apagada, e o sistema começou a tentar recolher do início. Não foi tentada uma nova tentativa.

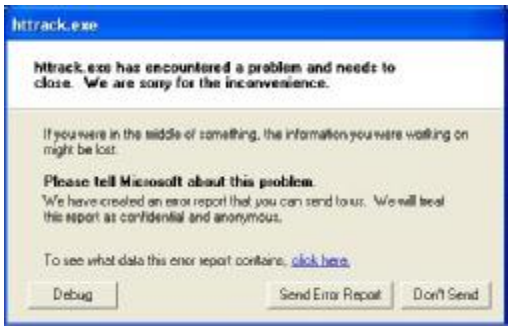

Figura 3 - Erro de execução do Httrack

Na recolha do sítio do departamento de Civil e Arquitectura (*civil.ist.utl.pt*) a existência de conteúdos dinâmicos ia criando directórios do tipo *civil.ist.utl.pt/pessoal/css/* ou *civil.ist.utl.pt/pessoal/img/* em que a última porção do endereço (css e img) era repetida indefinidamente, originando grandes quantidades de volume de tráfego e ficheiros desnecessários. Foi introduzido um filtro para evitar esta situação.

No sítio do Gire (gire.ist.utl.pt) há um menu lateral em *javascript* em que as ligações são varridas e os ficheiros descarregados, mas onde falha a navegação depois de recolhido porque a *applet* não é carregada.

Na página do centro de modelização de reservatórios petrolíferos, (*cmrp.ist.utl.pt/*) foi preciso iniciar a recolha a partir de uma página interior e não da inicial por causa dos *aliases* que impediam a navegação depois de o recurso estar recolhido.

4. Resultados

De acordo com os dados disponíveis, foram recolhidos cerca de 60 *gigabytes* distribuídos por cerca de 9500 recursos. Este valor não reflecte a quantidade de material disponível no espaço **.ist.utl.pt*, uma vez que ficaram por recolher com êxito os recursos relativos ao Fénix, ao departamento de Matemática, a subdomínios não detectados, utilizadores do alfa (*alfa.ist.utl.pt/~user*), e diversos servidores `ftp` existentes.

Certos conteúdos dinâmicos, como os que permitem múltiplas vistas sobre uma mesma página, originaram uma quantidade de dados superior à real. Alguns dos recursos recolhidos foram finalizados com um tamanho superior ao que está disponível em linha.

As recolhas efectuadas tiveram em conta as restrições impostas pelos `robots.txt`, pelo que determinadas páginas, como a página principal do IST, foram recolhidas sem a presença das folhas de estilo e determinadas imagens, pelo que o aspecto das páginas recolhidas não é o mesmo que as páginas originais. Não foi possível recolher recursos protegidos como o fórum da RNL, (*forum.rnl.ist.utl.pt*), que necessita de registo e autenticação, ou os documentos disponíveis no sítio do conselho directivo, por apenas estarem disponíveis dentro do domínio do IST.

A presença de conteúdos dinâmicos originou falhas na recolha de certos recursos. Por exemplo, na página do Núcleo de Física (*nfist.ist.utl.pt*) há galerias de imagem dinâmicas que não foram recolhidas. São ligações em *javascript* que apontam para imagens que, por causa disso, não foram recolhidas.

Durante a realização deste trabalho houve sítios que desapareceram e outros que mudaram de endereço. A tuna masculina, que estava alojada em *tuist.ist.utl.pt*, passou redireccionar a navegação e a alojar o seu conteúdo em *www.tuist.net*, ou seja, fora do domínio **.ist.utl.pt*, pelo que a página não foi recolhida. Mais: a tuna feminina, que está alojada em *tfist.ist.utl.pt* tem partes alojadas no sítio da tuna masculina que, por ter sido removida, deixaram de estar disponíveis no sítio da tuna feminina.

Houve ainda sítios que foram reformulados durante a realização deste trabalho, como é o caso do Instituto de Sistemas e Robótica, em que ao corrigir os erros da recolha que havia sido feita inicialmente, acabou por ser recolhida uma versão completamente nova do sítio.

Como seria de esperar, os sítios simples baseados em html foram recolhidos sem problemas. Pelo contrário, sítios com conteúdos em *flash* não são bem recolhidos, como é o caso do sítio da Secção Autónoma de Engenharia Naval (*http://www.mar.ist.utl.pt/*) cujos ficheiros até foram recolhidos mas onde depois falha a navegação no recurso final.

5. Análise Estatística

De acordo com os dados disponíveis foram recolhidos 59.617.591.413 *bytes* distribuídos por 911.790 ficheiros.

Na tabela 1 (página seguinte) podemos constatar que a maioria dos ficheiros eram do tipo html (asp e php estão incluídos), ocupando cerca de sete *gigabytes* e meio, e distribuídos por cerca de seiscentos e sessenta mil ficheiros. O segundo tipo de ficheiros mais comum é de ficheiros de texto, ocupando quase dez *gigabytes* em oitenta mil ficheiros.

As imagens (jpeg, gif e png) ocupam, todas juntas, cerca de nove *gigabytes* em cerca de cento e trinta mil ficheiros. Os ficheiros de áudio e vídeo (avi, mpeg, mp3) ocupam também grande espaço, com cerca de oito gigas em cerca de seiscentos ficheiros. Os ficheiros mp3 são maioritariamente provenientes do sítio da rádio interna do IST (*riist.ist.utl.pt*) e do seu fórum (*mega.ist.utl.pt/~riist/forum*) onde foram recolhidos ficheiros provenientes do podcast da rádio e mais de 13 horas de emissão da rádio em fluxo contínuo.

O tipo de ficheiros que ocupa mais espaço é do tipo arquivo tar que é, na sua esmagadora maioria, proveniente do recurso *ftp.ist.utl.pt*. Os restantes ficheiros de arquivo (zip, gzip) ocupam ainda cerca de três gigas dispersos por cerca de dois mil e quinhentos ficheiros.

Os documentos do tipo pdf e office (pdf, word, ps, pps) ocupam cerca de sete *gigabytes* reunidos em cerca de catorze mil ficheiros.

Os quase dois *gigabytes* correspondentes aos ficheiros do tipo `message/rfc822` são de mensagens de correio electrónico provenientes do arquivo das listas de correio do departamento de Matemática.

Verifica-se ainda um grande número de ficheiros *flash* e *javascript*: mais de mil e quinhentos ficheiros em duzentos *megabytes*. Os conteúdos em *flash* e *javascript* são muitas vezes responsáveis pelas recolhas mal sucedidas de recursos em linha, como também aconteceu na execução deste trabalho.

MIME	Ficheiros	Tamanho(B)
text/html	663.952	7.484.152.113
text/plain	80.484	9.770.570.809
image/jpeg	68.726	4.034.720.730
image/gif	54.207	414.465.071
image/png	12.234	598.273.373
application/pdf	10.868	4.961.218.488
application/x-tar	5.027	16.653.005.463
application/octet-stream	1.949	609.684.574
application/xml	1.612	9.882.371
text/css	1.427	5.013.026
application/zip	1.381	2.432.127.402
application/vnd.ms-powerpoint	1.246	1.237.643.604
application/msword	1.232	224.500.007
application/postscript	858	270.726.303
application/x-javascript	847	8.653.244
application/x-gzip	806	315.316.211
application/x-shockwave-flash	736	199.370.214
image/bmp	654	169.619.316
application/x-sh	444	351.816
text/xml	359	2.527.411
application/vnd.ms-excel	344	34.361.095
video/x-msvideo	336	3.633.887.268
model/vrml	289	7.745.358
audio/midi	219	5.357.383
audio/mpeg	217	3.932.694.585
application/x-perl	197	910.361
text/rtf	146	15.955.850
application/x-compress	110	20.726.039
image/x-icon	93	319.271
video/mpeg	89	432.797.980
chemical/x-pdb	75	1.760.888
audio/x-wav	69	11.523.744
application/x-zip-compressed	65	119.758.775
video/unknown	49	1.733.382
image/tiff	43	46.994.745
text/x-sh	40	104.431
application/xhtml+xml	30	432.845
video/quicktime	27	72.052.109
application/ogg	20	924.180
application/x-java-jnlp-file	17	27.225
application/mac-binhex40	16	10.883.923
audio/x-pn-realaudio-plugin	16	12.296.467
application/x-tcl	15	117.686
application/x-director	13	446.472
image/x-ms-bmp	12	273.760
text/javascript	12	118.103
application/x-msdos-program	10	12.533.880
application/x-java-vm	10	77.154

MIME	Ficheiros	Tamanho(B)
application/java-vm	10	72.610
audio/x-mpegurl	10	471
	10	122.483.912
application/vnd.rn-realmedia	9	1.907.551
audio/unknown	9	282.728
application/x-tex	8	559.566
application/x-httpd-php	8	60.877
message/rfc822	8	1.694.354.910
application/x-shar	7	10.226.314
application/atom+xml	6	69.945
application/x-texinfo	6	321.492
audio/x-pn-realaudio	6	951.514
application/x-troff	6	2.682
audio/basic	5	93.257
application/vnd.sun.xml.impress	5	1.568.115
video/x-ms-asf	4	1.746
text/vnd.wap.wml	4	1.329
application/x-msmetafile	4	24.582
application/xml-dtd	3	46.328
application/x-x509-ca-cert	3	5.656
application/rdf+xml	3	20.498
application/x-java-applet	3	13.935
application/txt%	3	2.292
text/x-server-parsed-html	3	4.710
application/x-ms-wmz	3	6.309
application/x-dvi	3	358.504
image/x-xbitmap	2	762
application/excel	2	1.285.632
application/x-gtar	2	20.480
application/msaccess	2	1.236.992
application/x-troff-me	2	9.130
application/x-stuffit	2	2.882.028
application/x-pdf	1	78.051
x-world/x-vrml	1	5.648
image/pjpeg	1	18.202
text/x-vcard	1	417
audio/mid	1	6.930
application/x-netcdf	1	4.250
image/svg+xml	1	48.555
application/binary	1	751.789
application/x-wais-source	1	164.856
application/wmz	1	293
application/xml+rss	1	3.093
application/x-java-archive	1	133.583
application/x-jar	1	17.528
text/x-tcl	1	36.995
application/x-chess-pgn	1	26.264.902

Tabela 1 - MIME Content Type, número e tamanho dos ficheiros recolhidos

6. Sugestões e Propostas de Melhoria

6.1. Severidade Alta

O facto de ocorrer um erro de execução no `Httrack` não devia ser motivo para o sistema apagar toda a recolha que já havia sido feita e recomeçar de novo. Por exemplo, no caso da recolha do departamento de Matemática, já haviam sido recolhidos cerca de oito *gigabytes* que deviam ser mantidos, uma vez que até foi uma tentativa de recolha que chegou mais longe que as anteriores. Podia, por exemplo, ser duplicada automaticamente a recolha e ser tentada novamente a duplicação e não a original.

Por vezes, surgem excepções no `Httrack`. Quando surge uma excepção no `Httrack`, é necessário fechar a janela que surgiu com o aviso da ocorrência dessa excepção para que o sistema continue a processar o que havia ficado em espera no *daemon* respectivo. O problema é que quando o *daemon* tem uma lista grande de acções para executar, ficam todas em espera, quando devia poder avançar. O sistema devia poder verificar automaticamente que um *daemon* não está a progredir e avançar para as acções seguintes. É como se não fosse preciso ter de desligar manualmente a janela de aviso da excepção. Se estiverem três *daemons* com uma carga grande em cada um deles e ocorrer um erro em cada um, todo o sistema fica parado, quando na verdade devia poder progredir. Quando se fecha uma janela de aviso de excepção do `Httrack`, a recolha é dada como concluída, como se nada tivesse acontecido, quando na verdade a recolha não foi completada com êxito. A recolha em vez de ficar com o estado de recolha bem sucedida, devia ficar com um estado que permitisse perceber que havia sido travada por uma excepção.

Não se deve confundir o erro de excepção com o erro de execução. Tanto quanto foi possível registar do comportamento do sistema, no erro de excepção a recolha é dada como concluída e os ficheiros recolhidos são mantidos. No erro de execução, muito mais raro, a recolha é dada como falhada, os ficheiros são apagados e é feita uma nova tentativa.

Quando eram adicionadas novas acções de recolha, verificou-se que essas acções eram muito frequentemente atribuídas a *daemons* ocupados em vez de serem atribuídas a *daemons* livres. Por exemplo, se estivessem dois *daemons* com acções em execução e duas ou três acções em espera, as novas acções em vez de serem atribuídas a um dos oito *daemons* livres, eram atribuídas aos dois *daemons* ocupados, fazendo com que apenas fossem tratadas depois de ser tratada toda a lista em espera. O sistema devia atribuir as novas acções a *daemons* livres. Por vezes, as acções que estavam a ser executadas eram recolhas de grandes dimensões com alguns *gigabytes* e era preciso esperar alguns dias para poder recolher as acções que havia sido adicionadas e que afinal só tinham uns *megabytes* para recolher, podendo ter sido despachadas logo na altura.

A gestão dos *daemons* de recolha podia ser melhorada de outras formas. Quando existem *daemons* livres, o sistema devia poder retirar acções em espera da lista de *daemons* ocupados e distribuí-las pelos *daemons* livres. Por vezes há dois ou três *daemons* com listas de espera enormes e os restantes *daemons* estão livres.

A robustez do sistema deve ser melhorada. O sistema deve estar preparado para gerir grandes quantidades de recursos, operando em simultâneo as recolhas e o cálculo de estatísticas. O cálculo de estatísticas deve, se possível, ser optimizado.

O `Httrack` deve ser melhorado de forma a funcionar com recursos como o Fénix.

6.2. Severidade Média

Detectaram-se erros na eliminação de recursos. Por razões não inteiramente compreendidas, determinados recursos que foram eliminados não conseguiram ser removidos com sucesso, pelo que permanecem na pasta Reciclagem, por remover.

Seria bom se ao calcular as estatísticas fosse possível ver se ainda falta muito ou não para completar os cálculos. Poderia ser feita, por exemplo, em relação ao número total de recursos e ao número de recursos já calculados, ou de outra forma que permitisse pelo menos saber o estado de evolução do cálculo de estatísticas em relação ao total. Da maneira actual apenas é possível saber se o sistema está a evoluir ou não.

A paragem de recolhas em acção deve melhorada. O sistema deve permitir que o utilizador mande parar uma recolha, podendo eventualmente retomá-la mais tarde. Por exemplo, se uma recolha for demasiado extensa o utilizador deve poder pará-la para, eventualmente, processar as restantes recolhas, que podem até ser mais pequenas, e continuar depois a primeira recolha. A paragem de uma recolha deve poder ser feita recolha a recolha e não apenas colecção a colecção.

Deve ser acabada a funcionalidade que permite a exportação de objectos.

6.3. Severidade Baixa

O ecrã do estado dos *daemons* de recolha poderia mostrar a velocidade a que os *daemons* estão a recolher informação.

Ao utilizar o sistema o utilizador apenas tem noção de avanço da recolha de um recurso através da análise dos registos de ligações (*logs* de "todos os *links*") onde vai sendo vista a progressão dos ficheiros descarregados. O utilizador devia poder ver a progressão de outra forma, por exemplo, a quantidade total de bytes e o número total de ficheiros descarregados. O utilizador só sabe quanto é que uma recolha já recolheu se fizer as contas manualmente a partir do *log* dos *links*. Em recolhas com grande número de ficheiros, esta tarefa manual torna-se muito difícil.

O sistema devia permitir extrair uma lista com os endereços dos recursos existentes. Desta forma seria fácil obter uma relação dos recursos existentes no sistema e respectivos endereços.

Quando se criam múltiplos recursos, e é preciso especificar a colecção, o *adicionar* (>) fica muito distante da lista de colecções e, se as estatísticas ainda não tiverem sido calculadas, o espaço intermédio aparece a branco e às vezes não se percebe a linha correspondente à colecção. Ainda neste ecrã, as *tips* de "seleccionar itens escolhidos" dizem "objecto com remoção agendada". Não faz sentido.

Na interface do utilizador, no ecrã de uma recolha, deve ser explicitado melhor o que é que expande para baixo e o que é ligação para uma página nova. Por exemplo, o aspecto da ligação "tipo de ficheiros" é idêntico ao do "*log* da recolha". No entanto, a primeira apenas expande informação enquanto a segunda abre uma nova janela. Podia, por exemplo, ser colocado um sinal gráfico a indicar que determinada informação apenas expande ali naquela região.

Quando se criam múltiplos recursos a partir de uma colecção, a colecção devia ser automaticamente seleccionada, para não ser preciso voltar a seleccionar. O facto de se poderem escolher várias colecções não deve obstar à selecção automática da colecção corrente.

No recurso da APAE (*http://aero.ist.utl.pt/~apae/*) o sistema desistiu da recolha por excesso de ligações mas não ficou indicada como uma recolha de grandes dimensões. Há situações em que se usam menus de navegação do lado esquerdo mas depois de recolhido falta uma barra nos endereços impossibilitando a navegação a partir desses menus, apesar de os ficheiros serem recolhidos, como aconteceu com a Associação dos Antigos Alunos (*http://aaa.ist.utl.pt/*). Estas ocorrências devem ser revistas.

As páginas da máquina principal que serve, por exemplo, a página do IST ou do Diferencial, remetem para uma página personalizada quando um ficheiro não é encontrado (um 404 personalizado). Nos *logs* que o `Httrack` produz, essas páginas são dadas como movidas e não como faltando, pelo que não indicam a página de onde são chamadas, impossibilitando saber que páginas contêm ligações para ficheiros inexistentes. Nos *logs* devia aparecer a página de onde o endereço foi chamado, mesmo que aparente ser uma página movida.

Deve ser criado, também, um manual de utilizador, visto que o sistema não é, por vezes, de imediata apreensão por parte de utilizadores menos habituados ao sistema e algumas das opções disponíveis são de dificuldade avançada.

7. Considerações Finais

Através da análise deste trabalho podemos concluir que a recolha de recursos da internet é uma tarefa lenta e que consome muitos recursos. A recolha bem sucedida requer constante vigilância dos processos de recolha e do sistema.

O sistema RECOLHA, associado ao `Httrack`, é um sistema que produz bons resultados e que tem ainda muita margem para melhorar. Há situações com as quais qualquer sistema tem dificuldade em lidar e o RECOLHA regista também, naturalmente, situações idênticas. Factores como a carga de recursos imposta ao sistema influenciam a capacidade de resposta do mesmo. O melhoramento do sistema deve ser continuado e a progressão do RECOLHA, que se verificou em constante evolução no decorrer deste trabalho, deve ser mantida.

Um agradecimento final ao apoio prestado pelo Professor José Borbinha e pelo Engenheiro João Luzio, sem os quais este trabalho não teria sido possível.

8. Anexos

No CD em anexo pode ser encontrada uma lista dos endereços identificados no espaço **.ist.utl.pt* e utilizados como recursos a recolher neste trabalho.