Caracterização da Informação WWW na RCCN

Maria João Nicolau, Joaquim Macedo e António Costa
Departamento de Informática
Universidade do Minho

(Versão Preliminar - Março de 1997)

Resumo

Discute-se um conjunto de medidas para a caracterização dos serviços de informação internet na RCCN (WWW, preferencialmente). Propõe-se um sistema para o registo cronológico dos valores desses parâmetros e a geração de gráficos com a respectiva variação no tempo e apresenta-se, a título ilustrativo, os valores dessas métricas para as várias instituições que integram a RCCN.

1 Introdução

É um dado adquirido que o WWW é bastante popular na Comunidade Académica em Portugal. O objectivo deste artigo é contribuir para *medir essa popularidade* e apresentar alguns números que contribuam para a sua melhor compreensão.

A popularidade pode ser medida com base em:

- quantidade de informação mantida,
- número de acessos a essa informação, e
- número de bytes transferidos

Não se pretende de forma alguma fazer uma abordagem exaustiva ao tema, mas fazer o levantamento de algumas medidas do espaço de informação WWW disponibilizado pelas organizações da RCCN e retirar algumas conclusões.

Do serviço WWW podem ser retiradas as mais variadas medidas que vão desde o tráfego http passando pela análise de *registos históricos* de servidores e clientes, até à análise da própria estrutura da informação.

Da análise do tráfego http pode ser inferida a quantidade e qualidade dos recursos de rede necessários e também a sua percentagem de utilização no tempo. É possível retirar informação mais rica se o tráfego for analisado a partir dos registo de históricos dum servidor *proxy*. Podem ser contabilizados aspectos relacionados com a carga de servidores e popularidade das páginas. As estatísticas do *proxy* WWW da RCCN[1] dão-nos uma ideia da informação que é possível extrair. .

Da análise dos registos históricos dos servidores é possível deduzir a carga dos servidores e também a popularidade das páginas mantidas. Um aspecto interessante que pode ser estudado é, por exemplo, os padrões de utilização típicos de um conjunto de páginas. Este estudo complementa o que é possível num servidor *proxy* dado que existe mais informação disponível. No entanto,alguns aspectos relacionados com a caracterização dos utilizadores só é conseguido com análise dos registos históricos dos clientes.

Algumas medidas só podem ser obtidas analisando as páginas em si[2]. São exemplos a consistência das referências, correcção da sintaxe HTML, etiquetas HTML utilizadas e também alguns dados disponibilizados pelas abordagens anteriores.

A indisponibilidade de alguns dos dados aqui referidos impedem que se façam análises em larga escala em qualquer ponto da rede. Adicionalmente, algumas das medidas são sensíveis em termos de confidencialidade dos indivíduos e das organizações.

Um outro aspecto importante prende-se com a geração de tráfego resultante da colecção de dados, quando esta é feita por *Robots*. Torna-se essencial explorar ao máximo as técnicas de monitorização passiva[3], no sentido da minimização do tráfego para o efeito.

Neste trabalho obtiveram-se algumas medidas com base na análise duma amostra das páginas WWW existentes na RCCN.

Relativamente a cada referência foi colectada a seguinte informação:

- URL - trata-se do URL absoluto associado à referência. A partir deste campo pode retirar-se:
  - o método de acesso,
  - o servidor,
  - a porta e
  - o tipo MIME deduzido a partir do extensão do ficheiro;

- URL onde é referenciada;
- Tamanho do ficheiro (em *bytes*);
- Tipo de conteúdo (tipo/subtipo MIME);
- Data da última modificação;
- Validade da referência: válida, inválida, inacessível (por *timeout* ou qualquer outro motivo);

Enquanto o URL onde é referenciada é deduzido do contexto de processamento, os atributos seguintes são conseguidos usando o método HEAD do protocolo HTTP.

Com excepção das referências http, a maior parte destes campos ou não está definida ou pode não fazer sentido.

Como o acesso às referências é feito pelo percurso das páginas, pode coleccionar-se informação adicional relativa a cada página, como por exemplo:

- número de imagens,
- etiquetas HTML utilizadas,
- consistência das referências,
- etc...

Na versão actual do software não é ainda colectada informação relativa às etiquetas.

A informação colectada permite obter o seguinte tipo de medidas a respeito das características das páginas, consolidadas por domínio ou para todo espaço de informação em estudo:

- tamanho médio por página,
- número médio de referências por página,
- número médio de referências válidas e inválidas por página,
- número médio de referências inacessíveis por página,
- número médio de imagens por página,
- número médio de referências a uma página,

- tempo médio desde a última modificação,
- percentagem de páginas por formato, e
- percentagem de páginas por método de acesso.

Foi seleccionado um conjunto de métricas a cujo valor se tentou dar uma interpretação qualitativa:

- *fiabilidade* — esta medida dá informação a respeito da fiabilidade [4] das referências:

  $$fiabilidade = 1 - \frac{r_i}{r_t}$$

  onde r_i é o número de referências inacessíveis e r_t o número total de referências.

  A fiabilidade para um domínio pode ser analisada sob duas perpectivas: a de fonte e a de destino. No primeiro caso interessa considerar referências de saída das páginas do domínio e no segundo as referências para páginas desse domínio.

- *visibilidade* — esta medida dá informação a respeito da visibilidade[5] duma entidade num determinado espaço de informação que é proporcional ao número de páginas em que uma entidade aparece.

  A fórmula seguinte é apresentada ainda com algumas reservas:

  $$visibilidade_e = \frac{r_e}{r_t}$$

  r_e é o número de referências feitas à entidade e r_t é o número total de referências a entidades do mesmo tipo feitas nesse espaço. A entidade pode ser um documento, um servidor ou uma organização.

  Por exemplo, podemos estar interessados na visibilidade mútua entre duas organizações.

- *localidade* — esta medida dá informação da localidade [4] das referências:

  $$localidade = 1 - \frac{r_e}{r_t}$$

  r_e é o número de referências externas e r_t o número de referências totais.

  O domínio considerado como local pode ser um servidor, uma organização ou um país. Desta forma, podem obter-se medidas de localidade relativas ao servidor, à organização ou ao país.

Parâmetro	Tam.	Ref. Int.	Ref. ext.	Ref. entrada	Ref saída	Ref. val	Ref. inv.	Ref ina	Ref Totais
Valor	5,4k	3,78	0,87	5	5	4,29	0,06	0,31	5

Figura 1: Valores médios por página

- *diversidade* — esta medida dá informação a respeito da diversidade existente a um atributo que possa assumir um conjunto (normalmente restrito de valores) no espaço de informação em análise.

  Como exemplo podemos apresentar o método de acesso, o tipo de dados MIME, a língua dos documentos e mesmo as portas de protocolo. Embora já tenha sido feito algum trabalho, não foi ainda encontrada a fórmula adequada.

2 Espaço de informação

A amostra que foi analisada neste trabalho é constituída por 16617 páginas com um total de 118541 referências disponibilizadas por 24 organizações da Rede Académica Portuguesa. O tamanho da amostra foi condicionada pelas limitações da plataforma em que as ferramentas de recolha foram instaladas, conectividade com os servidores a analisar e o tempo disponível . As páginas processadas são as dos níveis superiores da hierarquia de hipertexto das organizações que nos parecem mais relevantes.

3 Resultados obtidos

Obtiveram-se uma série de resultados que são apresentados nas tabelas e figuras seguintes.

Começa-se por um conjunto de valores médios por página na figura 1.

Os resultados correspondentes aos tipos MIME, apresentados na figuras 2 e 3, correspodem perfeitamente ao que se esperava: as páginas com conteúdo *text/html* e *image/gif* correspondem a grande percentagem das páginas existentes. O mesmo se pode dizer do método de acesso *HTTP* relativamente aos restantes, como se pode observar na figura 4. As portas de protocolo usadas nos métodos de acesso foram omitidas por se usarem, quase esmagadoramente, as portas por defeito.

Relativamente ao tamanho de página (figura 5) pode observar-se que a maior parte tem tamanhos entre os 1 e 5k, e que o intervalo 500bytes a 10k inclui a sua quase totalidade.

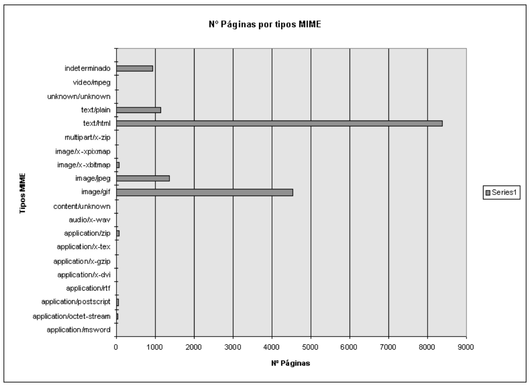

Figura 2: Páginas em pt por tipo/subtipo MIME

Na figura 6 estão apresentados, para o conjunto de domínios analisados, valores médios de número de referências internas e externas. Alguns deles, que correspondem precisamente aos domínios com maior número de páginas internas, foram analisados apenas os níveis superiores da hierarquia, pelo que o número de referências a páginas internas está bastante inflacionado.

Na figura 9 são dados os valores médios de referências de entrada e saída.

Inclui-se, apenas por curiosidade, o tamanho médio por página nos vários domínios (figura 8).

A seguir apresenta-se a consistência média por página incluindo número de URL inválidas, inacessíveis e válidas. Pode-se observar que as médias das referências inacessíveis é baixa.

Termina-se apresentando os valores médios das medidas com a semântica que foi apresentada anteriormente nas figuras 11, 12 e 13.

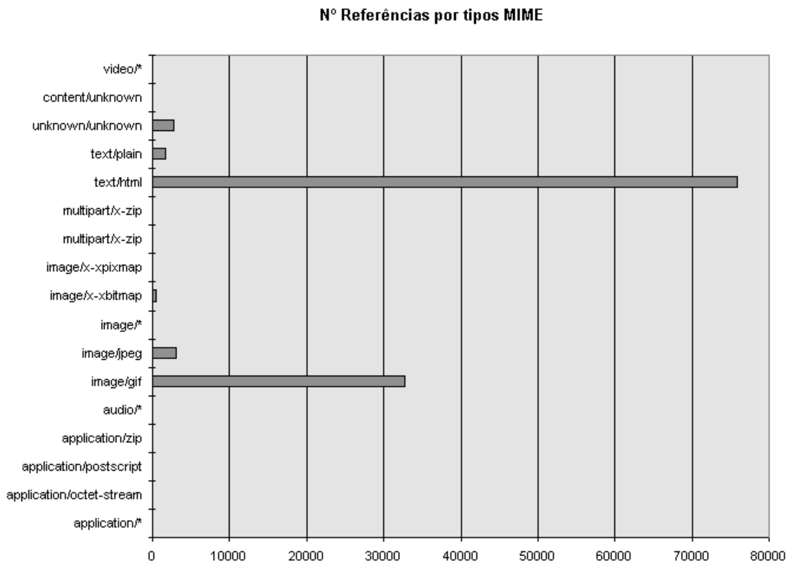

Figura 3: Referências em pt por tipo/subtipo MIME

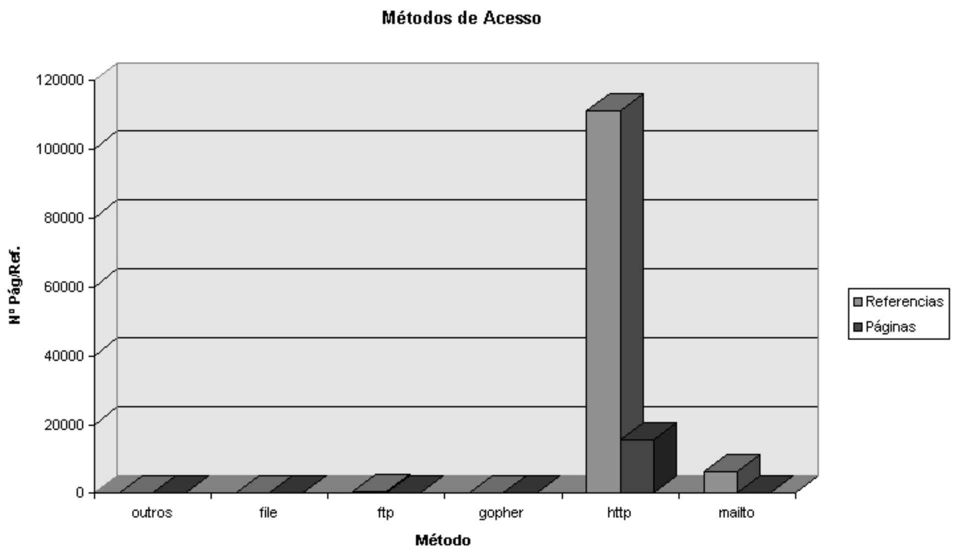

Figura 4: Distribuição de páginas e referências em pt por método de acesso

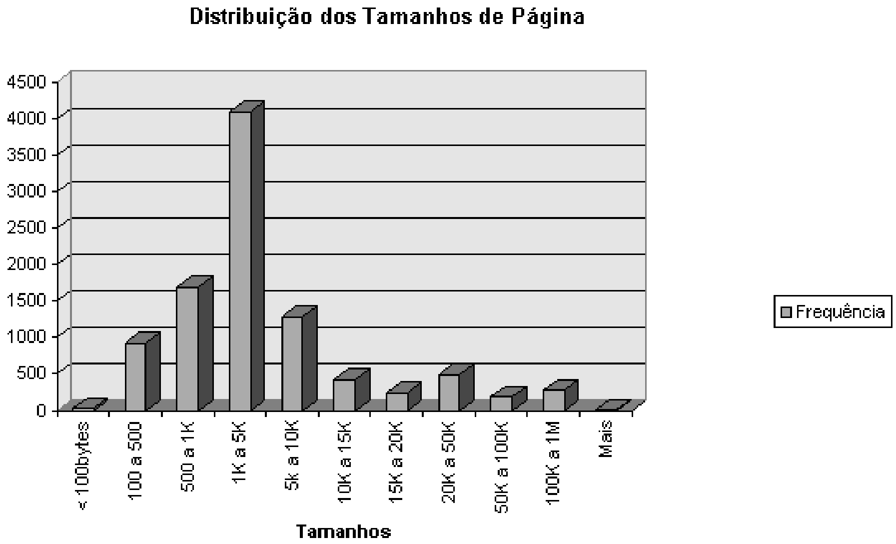

Figura 5: Distribuição de páginas por tamanho

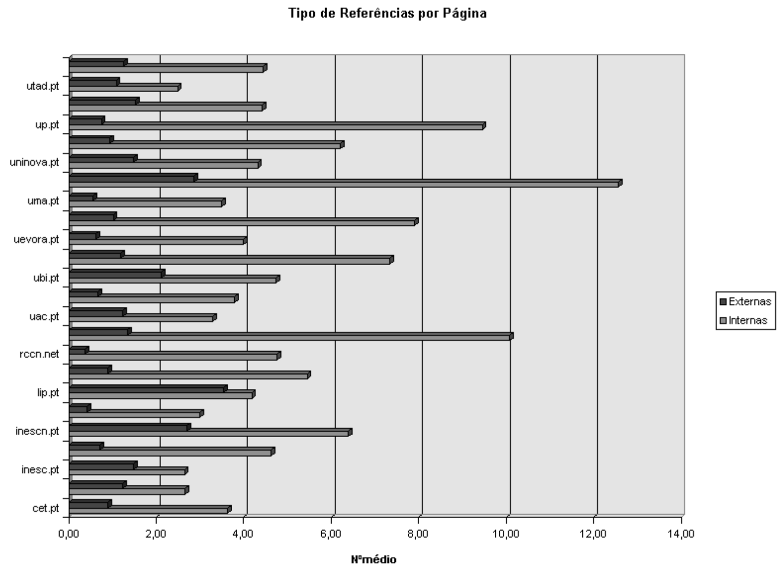

Figura 6: Número médio de referências internas e externas por página

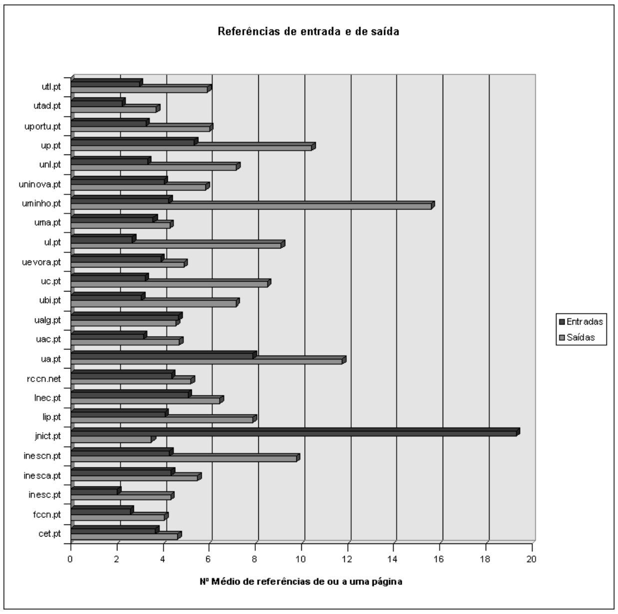

Figura 7: Tamanho médio por página em cada domínio

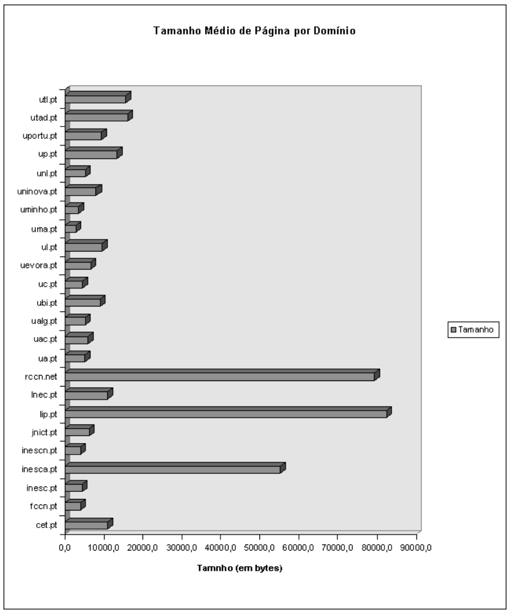

Figura 8: Tamanho médio por página em cada referência

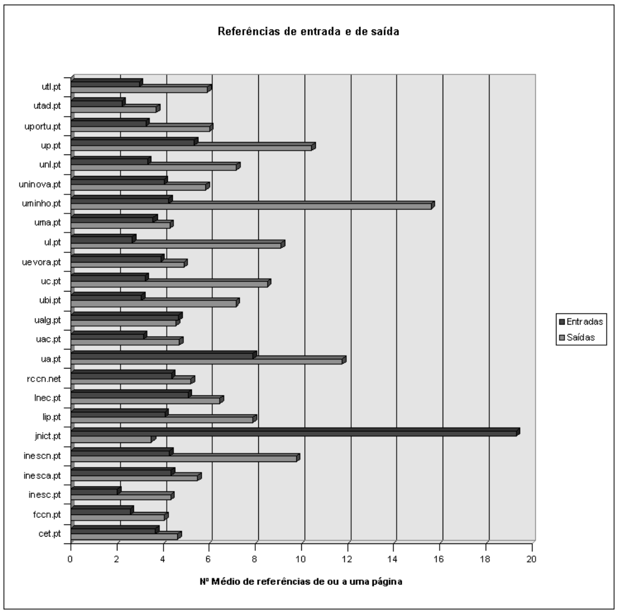

Figura 9: Número médio de referências de entrada e saída por página

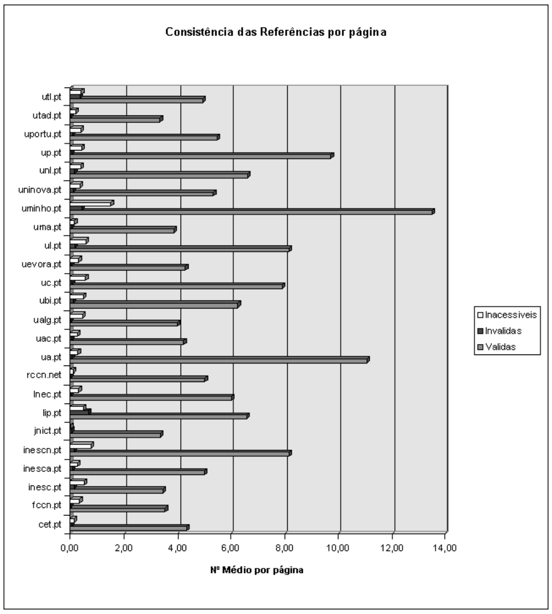

Figura 10: Consistência média das referências por página

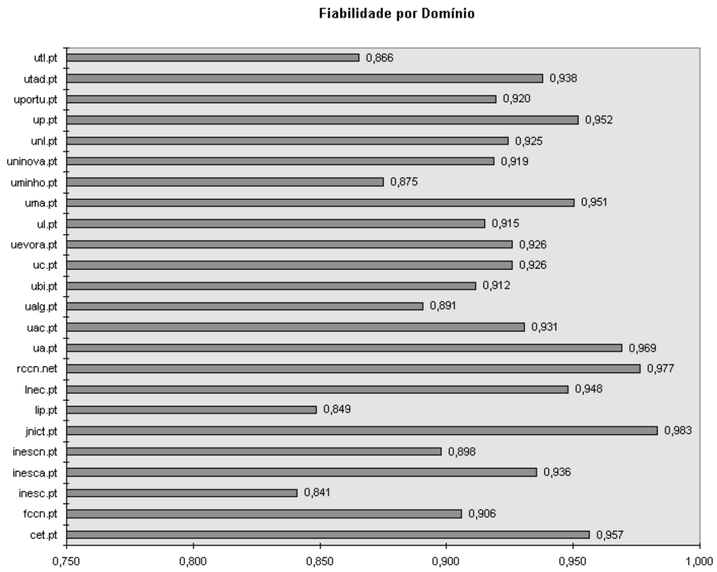

Figura 11: Fiabilidade média das referências por página

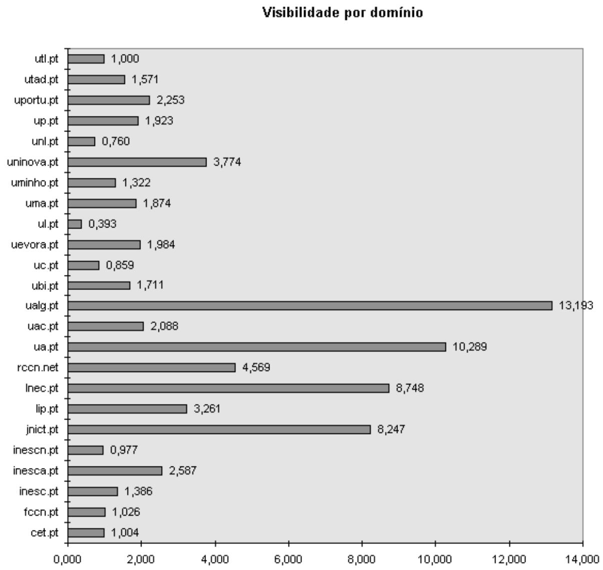

Figura 12: Visilidade média das referências por página

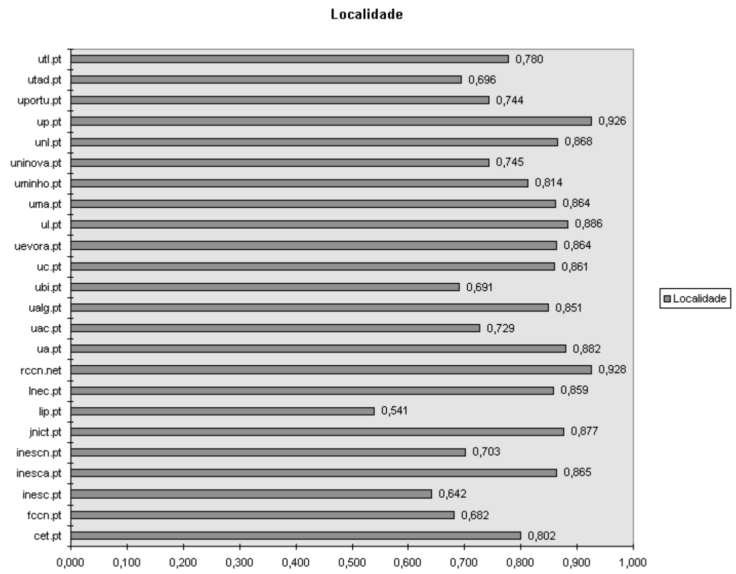

Figura 13: Localidade média das referências por página

4 Notas de concretização

O software utilizado (ver figura 14) está ainda numa versão preliminar e foi escrito em Java[6]. Neste instante ainda não tem a componente de monitorização passiva[1] [3] e usa os dados colectados por um robot. Os gráficos foram elaboradas usando uma folha de cálculo externa.

Referências

[1] RCCN. Serviço de Proxy. Disponível na Rede, 1997. http://www.fccn.pt/rccn/ servicos/proxy/stats/cache-stats.html.

[1] Ligação ao *proxy*

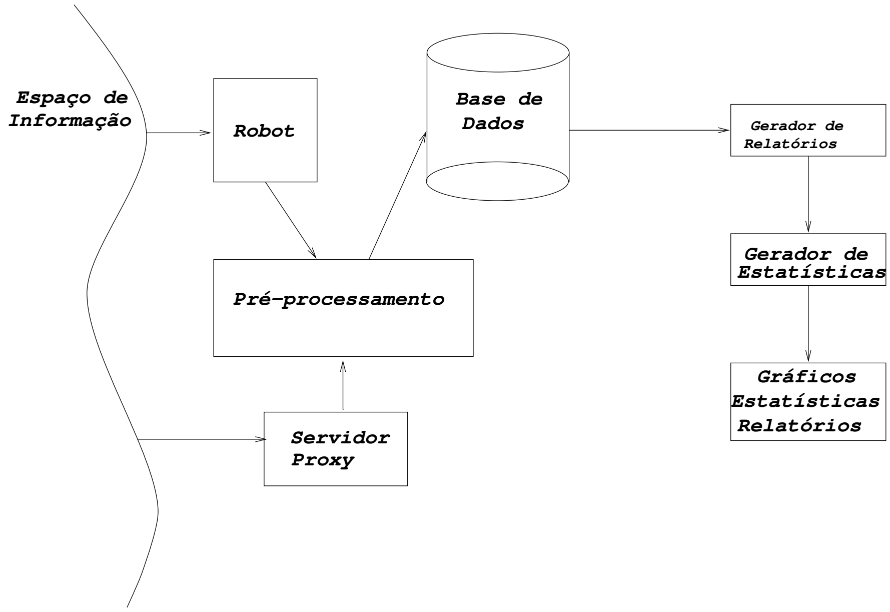

Figura 14: Componentes do Programa

[2] Allison Wodruff et all. An Investigation of Documents from the World Wide Web. In *Fifth Internacional World Wide Web Conference*, Paris,France, May 1996. University of California Berkeley.

[3] Paul Barker. Providing the X.500 Directory User with QOS Information. In *Computer Comunication Review - ACM SIGCOMM*, 1993.

[4] João Álvaro Carvalho Clara Cunha and Joaquim Macedo. *Caracterização dos Serviços de Informação da RCCN*. Technical report, Departamento de Informática, Universidade do Minho, Braga,Portugal, Janeiro 1996.

[5] Tim Bray. Measuring the Web. In *Fifth Internacional World Wide Web Conference*, Paris,France, May 1996. Open Text Corporation.

[6] David Flanagan. *Java in a Nutshel*. O'Reilly Associates, February 1996.