GeoCensus

Extracção de Informação Geográfica da *Web*

MACEDO, Joaquim e SANTOS, Maribel Yasmina

RESUMO

O sistema GeoCensus *tem como objectivo a extracção de informação geográfica da Web. Através da colecta exaustiva (para já apenas uma amostra significativa) de páginas Web existentes em Portugal, o sistema localiza geograficamente os servidores que hospedam as páginas colectadas. A localização geográfica das páginas pode ser realizada usando diferentes fontes de informação, nomeadamente a localização geográfica dos servidores Web que hospedam as páginas, os nomes geográficos extraídos do seu conteúdo e a localização geográfica das páginas adjacentes na topologia da Web. Estas fontes diferentes de evidência geográfica podem ser usadas, de forma combinada ou isolada, para localizar geograficamente as audiências das páginas (público alvo).*

A utilização de diferentes fontes de informação permite uma maior abrangência do sistema de localização geográfica, uma vez que uma grande parte das páginas não têm informação geográfica explícita ou apresentam poucos relacionamentos (nós de entrada e saída) na topologia da Web. Desta forma, a localização duma página pode ser determinada quer pela informação explícita que contém, quer pelas relações que estabelece na topologia Web (páginas que referencia ou em que é referenciada) e também pela localização geográfica do servidor Web em que está hospedada. Caso as diferentes fontes de evidência existam, a sua combinação pode permitir aumentar o eficácia do sistema.

Este artigo apresenta o sistema GeoCensus*, descrevendo os diversos componentes que integram a sua arquitectura e ainda as diferentes tecnologias utilizadas na sua concretização. São ainda sistematizadas as diversas heurísticas para cálculo estimado do âmbito geográfico das páginas. Para já, é apenas explorada a localização geográfica dos servidores que hospedam as páginas e as ligações entre as páginas estabelecidas pelas referências. São apresentados resultados ilustrativos com toda a informação necessária para o cálculo do âmbito geográfico das páginas.*

PALAVRAS-CHAVE: World Wide Web, Localização Geográfica, Exploração de Dados.

1 INTRODUÇÃO

A *World Wide Web* (Web) constitui uma importante fonte de informação para uma população cada vez maior de utilizadores em todo o Mundo. O seu crescimento exponencial, aliado à sua arquitectura distribuída e descentralizada, coloca inúmeros desafios à localização de informação relevante quer interrogando motores de busca quer navegando em catálogos ou classificados.

Uma técnica importante de redução do espaço de procura e estruturação de grandes volumes de informação é a referenciação geográfica de conteúdos Web. Esta técnica torna-se cada vez mais importante com o crescente acesso à Internet e à Web através de computadores nómadas.

O trabalho descrito neste artigo, enquadrado no projecto GeoCensus, tem como objectivo a extracção de informação geográfica da Web. A única fonte de informação utilizada é para já a informação publicamente acessível. A localização geográfica do conteúdo das páginas e do seu público alvo é conseguida pela combinação (ou utilização isolada) de diferentes fontes de evidência geográfica, nomeadamente a localização geográfica dos computadores hospedeiros, os identificadores geográficos explícitos ou implícitos no seu conteúdo e os relacionamentos estabelecidos na topologia Web.

O âmbito geográfico (que define uma audiência alvo) duma página no contexto da *Web* em Portugal, permite determinar se essa página pretende atingir um público ao nível internacional, caso em que por exemplo estará traduzida em várias línguas, a população de um país, região, cidade ou localidade. O âmbito geográfico obtido é constituído por um conjunto de identificadores, que correspondem a nós de diferentes níveis na hierarquia geográfica pré-definida. Por este facto, a combinação de evidências corresponde nos casos mais simples a operações de conjuntos (união e intersecção, por exemplo).

A hierarquia geográfica é mais profunda para Portugal (onde pode ir até à freguesia) mas no futuro cobrirá todos os países. Nas sub-árvores correspondentes a outros países, a profundidade será a menor possível. Embora a sub-árvore para Portugal seja a mais importante, importa localizar quer os conteúdos respeitantes a outros países, quer os servidores *Web* dos URLs (*Uniform Resource Locator*) referenciados ou replicados pelos conteúdos em Portugal.

O GeoCensus utiliza um sistema de posicionamento indirecto, referenciação espacial através de identificadores geográficos, na associação de uma dada página a determinada localização geográfica. O sistema de referenciação espacial através de identificadores geográficos foi concretizado recorrendo às pré-normas ISO (*International Standard Organisation*) para Informação Geográfica e Geomática [1].

A arquitectura do sistema GeoCensus integra três componentes: o Repositório de Dados, a Exploração de Dados e a Visualização de Resultados. O Repositório de Dados é o responsável pelo armazenamento dos dados geográficos e não geográficos utilizados pelos diversos módulos de extracção de informação geográfica que integram o componente de Exploração de Dados. O componente de Visualização de Resultados permite, com o auxilio de um Sistema de Informação Geográfica (SIG), visualizar e analisar a informação geográfica colectada.

Ao nível da concretização do sistema refere-se que as diversas bases de dados que o integram , estão a ser suportadas por sistemas gestores de bases de dados relacionais, recorrendo no caso do componente de Repositório de Dados ao ORACLE e no caso específico da base de dados cartográfica que integra o componente de Visualização de Resultados, ao *Microsoft Access*. O SIG utilizado é o *GeoMedia Web Map* v4.0 da *Intergraph Corporation*.

Que seja do conhecimento dos autores, trata-se do primeiro trabalho cujo objectivo é o uso combinado de diversas fontes de evidência geográfica, para localização exaustiva de conteúdos e endereços da Web em Portugal. Neste documento é apresentado o caso específico da localização geográfica de servidores.

Este artigo encontra-se organizado da seguinte forma: a secção 2 apresenta a arquitectura do sistema, descrevendo os seus principais componentes e ainda a interacção que existe entre os mesmos. A secção 3 descreve os diversos algoritmos utilizados pelo sistema. A secção seguinte, secção 4, sintetiza a concretização do sistema, salientando ainda as opções tecnológicas adoptadas na sua implementação. Na secção 5 são apresentados os resultados obtidos, procedendo-se à respectiva avaliação dos mesmo. Por último, a secção 6 sintetiza o trabalho desenvolvido, apresentando ainda algumas propostas de trabalho futuro.

2 ARQUITECTURA DO SISTEMA

O sistema GeoCensus (ver Figura 1) é constituído por três componentes principais: o Repositório de Dados, a Exploração de Dados e a Visualização de Resultados.

O Repositório de Dados integra a Base de Dados de Páginas (BDP), a Base de Dados Geográfica (BDG) e a Base de Dados de Palavras Chave (BDPC). A BDP é actualizada com a informação permanentemente recolhida pelo sistema GeoCensus. A sua integração com a BDG, que armazena os diversos identificadores geográficos considerados e ainda a hierarquia conceptual existente entre os mesmos, permite a exploração da BDP com o auxílio ao SIG. O conteúdo da BDPC é periodicamente actualizado, à medida que novos assuntos de interesse vão sendo identificados. Nesta base de dados diversas tabelas integram as palavras chave associadas a determinado domínio, às quais são associados os identificadores que permitem conhecer a sua localização.

O componente de Exploração de Dados integra um Robot para Descarga de Conteúdos (RDC), o Depósito de Conteúdos (DC), e os módulos de Localização Topológica, Localização de *hosts* e Análise de Conteúdos. O RDC é responsável pela descarga exaustiva de páginas da Web, as quais são descarregadas para o DC, com vista à posterior análise pelos três módulos considerados no sistema. O módulo de Análise de Conteúdos utiliza as tabelas armazenadas na BDPC no processo de identificação da localização indexada por determinado conteúdo.

A Figura 2 evidencia os diversos fluxos de dados que existem entre o componente de Repositório de Dados e os três módulos do componente de Exploração de Dados. O módulo de Localização Topológica e o módulo de Localização de *hosts* acedem à BDP (para leitura e escrita) e à BDG (apenas leitura), enquanto que o módulo de Análise de Conteúdos requer o acesso à BDP (para leitura e escrita), à BDG (apenas leitura) e à BDPC (apenas leitura).

O componente de Visualização de Resultados permite a estruturação e navegação através de diversos mapas temáticos, os quais são construídos com o auxílio de um SIG suportado por uma Base de Dados Cartográfica (BDC).

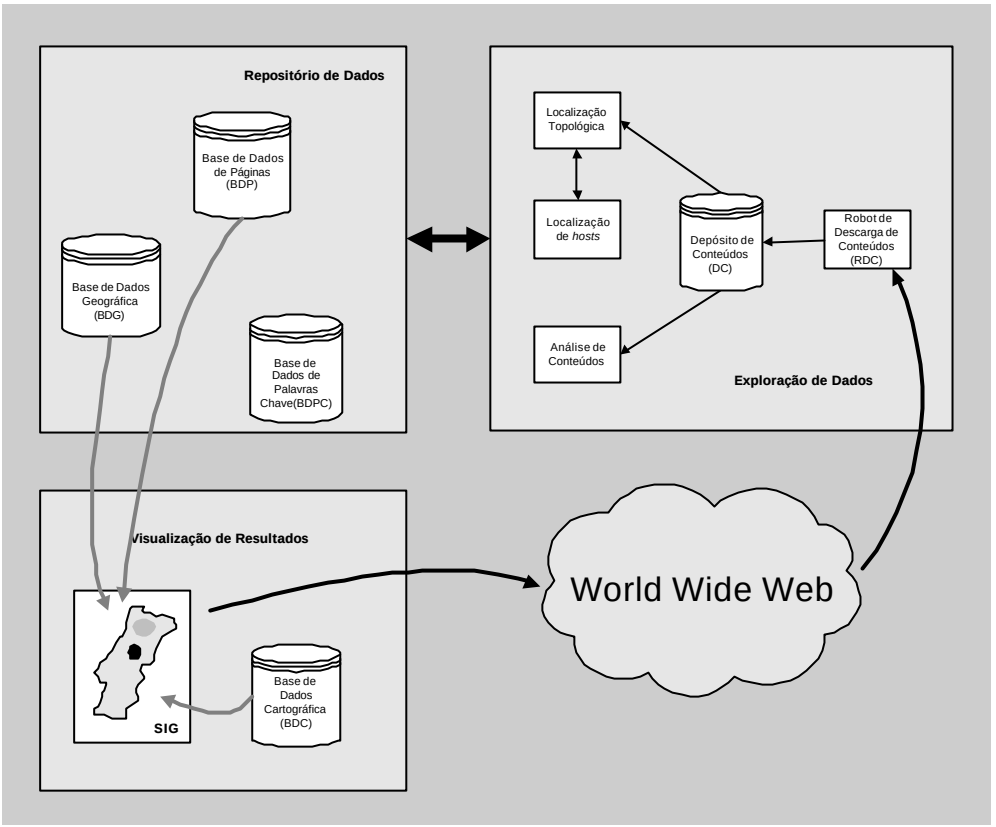

Figura 1 – Arquitectura do sistema GeoCensus

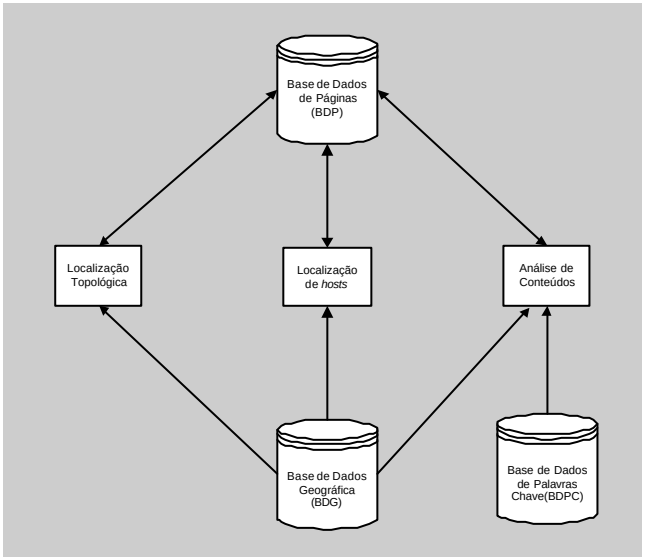

Figura 2 – Fluxos de dados

2.1 O Repositório de Dados

O Repositório de Dados é o responsável pelo armazenamento dos dados geográficos e não geográficos utilizados pelos diversos módulos considerados no componente de Exploração de Dados. De seguida é apresentada a estrutura das Bases de Dados (BD) que integram o Repositório de Dados.

2.1.1 Estrutura da BDP

O sistema GeoCensus constitui uma extensão ao sistema NetCensus [2]. O NetCensus tem como objectivo a concepção e concretização de um sistema automático (*hardware*, *software* e comunicações) que com um certo grau de supervisão permita a caracterização quantitativa e qualitativa do espaço *Web* em Portugal. De entre as várias entidades que caracterizam o modelo conceptual de dados do sistema NetCensus, a Figura 3 destaca as entidades que permitem a ligação ao sistema GeoCensus, e ainda, as entidades que possibilitam a especificação da componente geográfica (entidades sombreadas) dos servidores e das diversas páginas processadas. O registo de transação mantém toda a informação associada à interacção com o servidor remoto. Informação associada ao tempo de descarga, data de descarga, codificação de transferência, número de tentativas de descarga e informação do protocolo usado que é o HTTP. O registo URL tem a informação correspondente ao método de acesso e componentes do nome para além de apontar para o nome do servidor. O recurso tem informação relativa ao conteúdo como o seu tamanho, tipo, língua e alfabeto de caracteres. Os recursos estão acessíveis através de paginas HTML. Cada computador hospedeiro tem associado informação da sua localização. Do conteúdo da página podem ser extraídos identificadores com relevância geográfica e associar um conjunto de localizações. Conforme vai ser visto mais à frente pode-se calcular vários tipos de âmbitos geográficos que são colocados nos registos de âmbito.

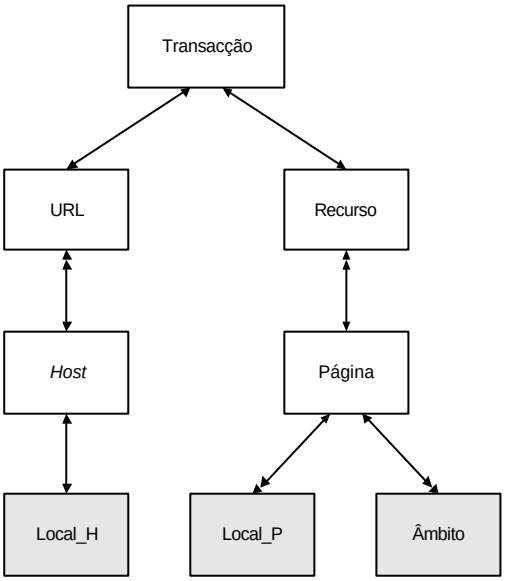

Figura 3 – Entidades da BDP

2.1.2 Estrutura da BDPC

A BDPC integra diversas tabelas que permitem associar às diversas instâncias de localização, consideradas pelo sistema de referenciação indirecto considerado pelo GeoCensus (o qual é apresentado com mais detalhe na subsecção 2.1.3), palavras chave que podem ser encontradas nas diversas páginas processadas pelo módulo de Análise de Conteúdos. Estas palavras chave, recolhidas manualmente, devem estar continuamente a ser processadas, permitindo aumentar a probabilidade de identificação do âmbito geográfico dos recursos analisados. A Figura 4 apresenta o modelo de dados da BDPC. Dentre as inúmeras tabelas de palavras chave que esta BD armazena destacam-se: Farmácias, Bombeiros, Câmaras Municipais, Hospitais, Barragens e Estádios, entre muitas outras.

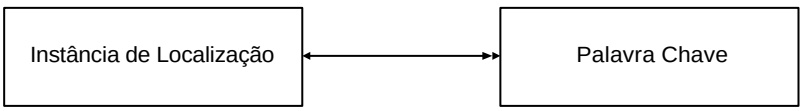

Figura 4 – Estrutura da BDPC

A Figura 5 apresenta um pequeno extracto do conteúdo da tabela correspondente ao Ensino Superior e ainda da tabela correspondente às Câmaras Municipais.

Ensino Superior

C_Concelho	Designacao	Tipologia
0603	Universidade de Coimbra	Ensino Superior Público Universitário
0105	Universidade de Aveiro	Ensino Superior Público Universitário
1312	Universidade do Porto	Ensino Superior Público Universitário
0303	Universidade do Minho	Ensino Superior Público Universitário
1714	Universidade de Trás-Os-Montes e Alto Douro	Ensino Superior Público Universitário
0503	Universidade da Beira Interior	Ensino Superior Público Universitário
0705	Universidade de Évora	Ensino Superior Público Universitário
0805	Universidade do Algarve	Ensino Superior Público Universitário
3103	Universidade da Madeira	Ensino Superior Público Universitário
4203	Universidade dos Açores	Ensino Superior Público Universitário

Câmaras Municipais

C_Concelho	Designacao	Presidente
0101	Câmara Municipal de Águeda	Manuel Castro Azevedo
0102	Câmara Municipal de Albergaria a Velha	João Agostinho Pinto Pereira
0103	Câmara Municipal de Anadia	Litério Augusto Marques
0104	Câmara Municipal de Arouca	José Armando de Pinho Oliveira
0105	Câmara Municipal de Aveiro	Alberto Afonso Souto Miranda
0106	Câmara Municipal de Castelo de Paiva	Paulo Ramalheira Teixeira
0107	Câmara Municipal de Espinho	José Barbosa Mota
0108	Câmara Municipal de Estarreja	José Eduardo A Valente Matos

Figura 5 – Extracto do conteúdo da BDPC

2.1.3 Estrutura da BDG

A BDG integra as diversas tabelas que permitem a caracterização do Sistema de Referenciação Espacial por Identificadores Geográficos utilizado pelo sistema GeoCensus. Este sistema considera as directivas especificadas pelo comité técnico TC 211 da ISO. A interpretação das recomendações especificadas no documento N 1172 [1] conduziu ao esquema relacional apresentado na Figura 6.

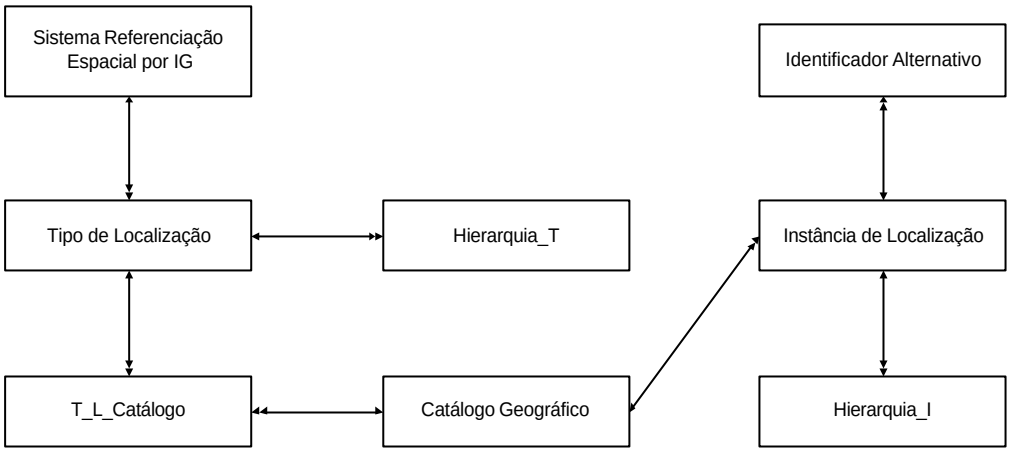

Figura 6 – Estrutura da BDG

A Figura 7 apresenta um pequeno excerto das tabelas Tipos de Localização do Catálogo (T_L_Catálogo na Figura 6) e da tabela de Hierarquia das Instâncias (Hierarquia_I na Figura 6) que integram a BDG.

Tipos_localização_Catálogo	
Nome_Tipo	**ID_Gazetter**
Distritos	1
Concelhos	1
Freguesias	1
Localidades	1
Areas_Metropolitanas	1
Nut_I	1
Nut_II	1
Nut_III	1

Hierarquia_instâncias	
Id_geográfico	**id_superior_geográfico**
A01	
A02	
A03	
B01	A01
B02	A01
B03	A01
B04	A01
B05	A01
B11	A02
B12	A03
101	B01
102	B01
103	B01

Figura 7 – Excerto do conteúdo da BDG

2.2 A Exploração de Dados

De seguida, são apresentados os módulos que interagem com a Web extraindo a informação necessária para o carregamento da BDP e sua posterior análise para a identificação da informação geográfica associada.

2.2.1 A Descarga de Conteúdos

O RDC é responsável pela colecta de dados com base em visitas que efectua à Web. É seguido o algoritmo convencional. Com base nos URLs semente que são colocados na fila de pendentes, descarrega-se o conteúdo de cada página. Se o conteúdo não tiver sido ainda processado, extraem-se os URLs que são também colocados na fila de pendentes. Para ser colocado na fila dos pendentes o URL tem que ainda não ter sido descarregado e tem que satisfazer um conjunto de filtros.

Um subconjunto seleccionado dos conteúdos descarregados (por exemplo os ficheiros HTML - *HyperText Markup Language*) são armazenados em disco local de forma estruturada. A informação associada a cada transação é mantida num registo XML (*eXtensible Markup Language*). O registo para além do apontador para o ficheiro fonte no DC mantém meta-informação sobre o conteúdo e sobre a operação de descarga.

```
<TRAN Url="www.uminho.pt" FetchDate="Tue Oct 23 19:17:10 WEST 2001"
DocumentFingerPrint="aa2d8c5bb9b41801" Length="3128">

<HTTPDATA HTTPVersion="http/1.0" HttpStatusCode="200" NumberOfCookies="0">
<HEADER key="Content-type" info="text/html"/>
<HEADER key="Transfer-Encoding" info="chunked"/>
<HEADER key="Server" info="Netscape-FastTrack/2.01"/>
<HEADER key="Contet-length" info="236"/>
</HEADERS>
</HTTPDATA>

<DNSDATA>
<IP ip="193.136.20.10"/>
<IP ip="193.136.20.12"/>
<IP ip="193.136.20.13"/>
</DNSDATA>

<HTMLDATA>
<Title> Universidade do Minho </Title>
<Description> Homepage da UM </Description>
<Author> Alguém </Author>
<Generator> Mozilla/3.0Gold (X11; I; OSF1 V4.0 alpha) [Netscape] </Generator>
<HTMLVersion> HTML PUBLIC -//W3C//DTD HTML 3.2//EN </HTMLVersion>
<COUNTS>
<COUNT TAG="Applet" COUNT="5"/>
<COUNT TAG="Area" COUNT="0"/>
<COUNT TAG="Caption" COUNT="2"/>
<COUNT TAG="Script" COUNT="0"/>
<COUNT TAG="Img" COUNT="7"/>
</COUNTS>
</HTMLDATA>

<LINKS>
<Link UrlDestiny="http://marco.uminho.pt:80/mc-icons/unknown.gif" Pos="1000"
TagHTML="img" AttrHTML="src"/>
<Link UrlDestiny="http://marco.uminho.pt:80/mc-icons/unknown2.gif" Pos="1030"
TagHTML="img" AttrHTML="src"/>
</LINKS>

<CONTENT Path="/usr1/save/fd00000/r00000"/>
</TRAN>
```

O conjunto de registos XML e conteúdos associados constitui o DC. O DC é posteriormente objecto dos mais diversos tipos de processamento. Para além dos processamentos relevantes para o GeoCensus, podem existir processamentos para outras aplicações como motores de busca, catálogos, arquivo e sistemas de obtenção de estatísticas de conteúdo, etc...

Um dos processamentos realizados é o carregamento da BDP, em que é mantida meta-informação sobre as páginas e respectiva relação topológica. A BDP é posteriormente enriquecida com informação de localização geográfica produzida pelos módulos específicos do GeoCensus.

2.2.2 Os módulos de exploração

Os três módulos de exploração dos dados considerados no sistema GeoCensus permitem a localização geográfica de *hosts*, a localização geográfica de páginas atendendo à topologia *Web* e a localização geográfica das páginas atendendo ao conteúdo das mesmas (estes dois últimos permitem a identificação do âmbito geográfico de um dado recurso *Web*).

2.2.2.1 Localização geográfica de *hosts*

Um dos módulos mais importantes do GeoCensus é o que se encarrega da localização geográfica dos *hosts* que hospedam os conteúdos identificados pelos URLs extraídos das páginas processadas.

Para a localização geográfica dos *hosts* é usado um algoritmo com vários passos. Cada um desses passos utiliza uma heurística baseada numa determinada fonte de informação. O primeiro passo é verificar se a localização geográfica dessa máquina já foi determinada em corridas anteriores, caso em que terá um registo com informação completa num ficheiro de *cache*. Seguidamente, é verificado se o DNS (*Domain Name System*) tem um registo do tipo LOC (Resource Record (RR) de Localização Geográfica). Se não existir, o passo é repetido para o campo de nível superior do nome. Os nomes correspondentes a máquinas de fornecedores de serviços IP (*Internet Protocol*, por exemplo Telepac, RCTS, etc.) normalmente têm informação de localização no seu nome.

Existem diversas bases de dados que são também consultadas no processo de localização, as quais incluem informação sobre a localização de *hosts*, organizações, cidades e aeroportos. Essas bases de dados são distribuídas com o GTRACE [3], o pacote de *software* usado como ponto de partida.

De seguida é apresentado um pequeno exemplo (ver Figura 8) do modo de funcionamento do módulo de localização geográfica de *hosts*. O ficheiro de entrada no módulo integra o URL do *host* a localizar e o seu respectivo endereço IP. Como resultado do processo de identificação da localização obtém-se um ficheiro de saída com os seguintes atributos: URL do *host*, endereço IP do *host*, latitude, longitude, cidade, estado e país.

Ficheiro de Entrada

```
www.ave.dee.isep.ipp.pt, 195.23.79.205
www.iplei.pt, 193.137.239.229
alma.dei.uc.pt, 193.137.203.233
www.ipcb.pt, 212.55.157.161
www.cm-gaia.pt, 212.55.149.3
```

Ficheiro de Saída

```
(www.ave.dee.isep.ipp.pt, 195.23.79.205,41.17,-8.58,Porto,,PT)
(www.iplei.pt, 193.137.239.229,39.75,-8.8,Leiria,,PT)
(alma.dei.uc.pt, 193.137.203.233,40.2,-8.42,Coimbra,,PT)
(www.ipcb.pt, 212.55.157.161,39.82,-7.5,Castelo Branco,,PT)
(www.cm-gaia.pt, 212.55.149.3,41.13,-8.62,Vila Nova De Gaia,,PT)
```

Figura 8 – Módulo de localização geográfica de *hosts*

2.2.2.2 Extracção de informação dos conteúdos

O conteúdo das páginas é processado para identificação de palavras chaves que possam funcionar como identificadores com relevância geográfica. Interessa extrair nomes de localidades, endereços postais, indicativos de número de telefones, nomes de autarcas, monumentos, estádios ou outras entidades ou pessoas às quais seja possível associar inequivocamente uma determinada zona geográfica ou localidade.

2.2.2.3 Identificação do âmbito geográfico

A disponibilização de conteúdos na Web pressupõe a existência de um público alvo, ao qual se destinam os conteúdos. A audiência de um conteúdo pode ser de âmbito restrito (limitado a determinada zona geográfica) ou ter uma abrangência alargada caso o mesmo seja relevante por exemplo a nível nacional. Neste trabalho utiliza-se a designação âmbito geográfico para explicitar a distribuição geográfica do público alvo de determinado conteúdo Web.

Adoptando a definição estabelecida por Ding *et al.* [4], o ***âmbito geográfico*** *que um recurso Web* w *referencia é a área geográfica que o criador de* w *pretende atingir*.

Na identificação do âmbito geográfico é considerada a hierarquia geográfica definida no sistema de identificadores geográficos estruturado anteriormente. A identificação do âmbito geográfico pode ser realizada por localização através da topologia *Web* (Localização Topológica) ou por Análise de Conteúdos. Estes dois módulos são detalhadamente descritos na secção 3 através da apresentação dos algoritmos que os implementam.

2.3 A Visualização de Resultados

O componente de Visualização de Resultados disponibilizará ao utilizador um conjunto pré-definido de mapas, os quais caracterizam a informação processada pelo sistema GeoCensus. Os mapas, a disponibilizar através de uma interface apropriada, permitem conhecer a:

1. distribuição geográfica de *hosts*. Esta visualização apresenta duas vertentes:

   - a distribuição quantitativa de todos os *hosts* contabilizados no sistema. Esta contabilização possibilita o agrupamento da informação por região, permitindo posteriormente a construção do respectivo mapa temático.
   - a identificação da localização de um dado *host*.

2. distribuição geográfica de conteúdos *Web* (URLs). Neste caso específico é possível:

   - quantificar, por região, todos os conteúdos catalogados.
   - dada uma região, distrito, quantificar os conteúdos por concelho.

3. localização geográfica de um conjunto de conteúdos. Para os conteúdos assinalados é ainda permitida a navegação através dos *links* integrados nos mesmos.

4. identificação do(s) *host*(*s*) em que reside(m) um dado conteúdo *Web* (URL). Neste caso específico, e a partir de um dado conteúdo, serão explicitadas graficamente todas as regiões em que residem *hosts* que alojam as páginas referenciadas por esse conteúdo.

5. localização geográfica dos URLs (não nacionais) referenciados por páginas HTML em Portugal.

Neste componente serão ainda utilizadas diversas operações de análise espacial (*overlap*, *meet*, *contain*, *touch*,…) disponibilizadas pelo SIG adoptado, para combinar os diferentes níveis de informação enumerados anteriormente.

3 ALGORITMOS UTILIZADOS PELO GEOCENSUS

Nesta secção são apresentados os algoritmos utilizados pelos módulos de Localização Topológica e de Análise de Conteúdos para identificar o âmbito geográfico de um dado recurso *Web*.

3.1 Por localização através da topologia Web

A identificação do âmbito geográfico através da verificação da topologia Web passa por verificar as diversas ligações existentes entre os recursos Web. Um recurso com âmbito nacional será apontado por diversos *links* distribuídos por páginas ao longo de todo o país, enquanto que uma página com âmbito local terá uma distribuição dos *links* mais restrita. Determinar o âmbito geográfico passa então por verificar a distribuição geográfica das ligações que existem para determinado recurso. Uma localização `l` integra o âmbito geográfico de um recurso `w` se verificar as seguintes condições:

1. uma parte significativa das páginas contidas em `l` apresentam ligações para `w`.
2. as páginas existentes em `l` que apresentam ligações para `w` deverão possuir uma distribuição uniforme ao longo de `l`.

Desta forma, o âmbito geográfico de um recurso `w` pode ser estimado pela identificação de um conjunto de localizações candidatas `l` que satisfazem as condições acima referidas. O processo de identificação das localizações candidatas passa pela determinação de duas variáveis: o `Interesse` de `w` em `l` e a `Uniformidade` do interesse por `w` ao longo de `l`.

A determinação do `Interesse` baseia-se na premissa de que uma localização `l`, que integra o âmbito geográfico de um recurso `w`, deve evidenciar um interesse relativamente alto por `w` ao longo das suas páginas [4]. Na prática significa que grande parte das páginas contidas em `l` devem conter ligações para `w`. O interesse é determinado através de `Interesse(w,l)= Ligações(w,l)/Páginas(l)`, em que `Ligações(w,l)` identifica o número de páginas inseridas na localização `l` que apresentam uma ligação para `w`, e `Páginas(l)` representa o número total de páginas Web existentes em `l`.

Para que uma localização `l` esteja inserida no âmbito geográfico de um recurso `w` é necessário que o interesse por esse recurso esteja uniformemente distribuído ao longo de `l`. A `Uniformidade(w,l)` é elevada quando

$Interesse(w,l_i) \sim Interesse(w,l_j)$ para todas as localizações l_i, l_j que são sub-localizações de l na hierarquia geográfica definida. O valor máximo que a Uniformidade pode adquirir é 1 (um), e tal será verificado em duas situações particulares. No caso de:

- l representar uma *folha* na hierarquia geográfica definida, pelo que o interesse por w ao longo de l é completamente uniforme. No caso da hierarquia conceptual definida para Portugal, as localizações representadas nas *folhas* da árvore correspondem às diversas freguesias do país.

- $Interesse(w,l)$ ser igual a 0 (zero), o que indica a inexistência de interesse por um determinado recurso ao longo de uma dada localização l.

A determinação da Uniformidade pode ser efectuada através de três formas distintas [4], para os casos em que l não represente uma folha na hierarquia geográfica e o $Interesse(w,l)>0$. Em ambas as definições $l_1, \ldots, l_n$ representam sub-localizações de l. Em ambas, para cada l é determinado: i) o vector $Páginas=(p_1, \ldots, p_n)$ que armazena o número de páginas $p_i=Páginas(l_i)$ de cada l_i de l (para $i=1, \ldots, n$); ii) o vector $Ligações=(h_1, \ldots, h_n)$, que lista o número de páginas $h_i=Ligações(w,l_i)$ que apresentam uma ligação para w na localização l_i (para $i=1, \ldots, n$); e ainda iii) o vector $Interesse=(r_1, \ldots, r_n)$ que identifica o interesse $r_i=Interesse(w,l_i)$ por w em cada sub-localização l_i de l.

De seguida são apresentadas as três definições que permitem o cálculo da Uniformidade: Modelo do Espaço Vectorial, Erro Relativo e Entropia.

Modelo do Espaço Vectorial

Baseada no modelo *do espaço vectorial* da área de *busca de informação*, permite determinar a similaridade existente entre os vectores Páginas e Ligações através do cálculo do coseno que os caracteriza. Se a quantidade de páginas com ligações para w é geralmente constante ao longo de todas as l_i ($i=1, \ldots, n$) que integram l, então o coseno do ângulo que caracteriza estes dois vectores será muito próximo de 1.

$$Uniformidade(w,l) = \frac{\sum_{i-1}^{n} p_i \times h_i}{\sqrt{\sum_{i-1}^{n} p_i^2} \times \sqrt{\sum_{i-1}^{n} h_i^2}}$$

Erro Relativo

Considerando que

$$R = (\sum_{i-1}^{n} h_i)/(\sum_{i-1}^{n} p_i)$$

então $r_i=R$ (para $i=1, \ldots, n$) se a distribuição do interesse por w é uniforme ao longo de l. Para medir o desvio que existe na uniformidade, calcula-se quanto cada r_i se distancia do valor ideal R. O valor da uniformidade é assim obtido no cálculo do erro relativo de cada l_i em relação a R.

$$Uniformidade(w,l) = \frac{1}{1 + \frac{1}{\sum_{i-1}^{n} p_i} \sum_{i-1}^{n} p_i \times \frac{|R - r_i|}{R}}$$

Entropia

Esta definição é baseada na noção de entropia da Teoria da Informação. Assume que existe uma fonte de informação associada com um recurso w e uma localização geográfica l. A fonte de informação gera símbolos que representam as diferentes sub-localizações de l, $l_1, \ldots, l_n$, através da execução consecutiva de três passos:

1. Selecção aleatória de um l_i;

2. Selecção aleatória de uma página localizada em l_i;

3. Geração de um símbolo que represente l_i, se a página seleccionada tem uma ligação para w.

Quando r_i=Interesse(w,l_i) é uniforme ao longo de todas as sub-localizações de l, a fonte de informação atingirá a entropia máxima na localização l, que é dada por log n. A uniformidade é calculada, utilizando esta definição para localizações geográficas com diferentes sub-localizações, através de:

$$\text{Uniformidade}(w,l)=\frac{-\sum_{i-1}^{n}\frac{r_i}{\sum_{j-1}^{n} r_j}\times\log\left(\frac{r_i}{\sum_{j-1}^{n} r_j}\right)}{\log n}$$

Âmbito Geográfico Estimado (AGE)

A determinação do AGE passa pela identificação dos âmbitos candidatos, atendendo ao Interesse e à Uniformidade calculados anteriormente. O Âmbito Geográfico Candidato (AGC) de um recurso w é constituído por um conjunto de nodos da hierarquia geográfica. Uma localização l pertence ao AGC(w) se satisfaz, dado um limiar t_c, as seguintes condições:

1. Uniformidade(w,l)= t_c;
2. Para cada l', superior hierárquico de l, Uniformidade(w,l')< t_c.

O AGC integra assim localizações com um interesse uniforme em w. Contudo, as condições apresentadas permitem que esse interesse possa ser reduzido, pelo que os âmbitos candidatos têm de ser "cortados" por forma a apenas incluírem, no AGE de um dado recurso, localizações com um interesse satisfatório.

O AGE de um recurso w, AGE(w), e que consiste num conjunto de localizações obtidas a partir do AGC(w), é determinado utilizando uma das seguintes estratégias de "corte":

- Primeiros k. Dado um número inteiro k, o AGE(w) consiste nas primeiras k localizações do AGC(w), por ordem decrescente do seu interesse.
- Limite Absoluto. Dado um limite t_e, AGE(w)={l ? AGC(w)|Interesse(w,l)= t_e}.
- Limite Relativo. Dada uma percentagem p, AGE(w)={l ? AGC(w)|Interesse(w,l)= max_Interesse(w) x p}, tal que max_Interesse(w)= max{Interesse(w,l)|l ? AGC(w)}.

3.2 Por análise de conteúdos

A identificação do âmbito geográfico de um recurso w, atendendo ao conteúdo do mesmo, passa por verificar a distribuição das localizações referenciadas no recurso. Para uma localização l integrar o âmbito geográfico de um recurso w deve satisfazer as seguintes condições:

1. Uma parte significativa das localizações mencionadas (directa ou indirectamente) em w referenciam l ou uma sub-localização de l.
2. As localizações referenciadas em w distribuem-se uniformemente ao longo de l.

Determinar se uma localização l pertence ao âmbito geográfico de um recurso w passa por calcular a variável Localizações(w) que representa o número de localizações referenciadas ao longo do texto que integra w, e ainda a variável Referências(w,l) que indica o número de referências a l encontradas em w. O interesse, atendendo ao conteúdo, é dado por Interesse(w,l)=Referências(w,l)/Localizações(l).

A uniformidade é neste caso calculada recorrendo aos vectores: i) Localizações=(p_1,…,p_n) em que cada localização armazena o mesmo valor p_i=Localizações(w) (para i=1,…,n), que representa o número de localizações referenciadas no texto que integra w; ii) Referências=(h_1,…,h_n) que lista o número de referências para cada sub-

localização l_i (para $i=1,\ldots,n$) no texto de w; e ainda iii) $Interesse=(r_1,\ldots,r_n)$ que identifica o interesse $r_i=Interesse(w,l_i)$ de w em cada sub-localização de l.

Estes três vectores são utilizados na determinação da uniformidade, tal como definida anteriormente, o que permite calcular e estimar o âmbito geográfico de um dado recurso w, baseado no seu conteúdo.

4 CONCRETIZAÇÃO DO SISTEMA

Esta secção tem como objectivo a descrição pormenorizada dos aspectos relacionados com a concretização do sistema GeoCensus. Os módulos de *software* aqui descritos foram baseados sempre que possível em módulos já existentes (concretizados por terceiros ou no âmbito de outros projectos), com funcionalidades acrescidas ou modificadas para este trabalho. Destaca-se a utilização do *GeoMedia Web Map* [5] para visualização de resultados, o sistema de gestão de bases de dados ORACLE no Repositório de Dados, e robot Larbin [6] e o GTrace [3] no componente de Exploração de Dados.

4.1 O Repositório de Dados

O Repositório de Dados é concretizado com base em programas em Java para carregamento e interrogação de BDs relacionais. Os esquemas das várias BD apresentadas, nomeadamente a BDP (Base de Dados de Páginas), a BDPC (Base de Dados de Palavras Chave) e a BDG (Base de Dados Geográfica) são concretizadas como tabelas no sistema gestor de bases de dados ORACLE.

Os problemas que se colocam principalmente na BDP são problemas de escala devido ao grande número de páginas e referências existentes no espaço Web Português. Num amostra em que foi colectada apenas uma porção da Web foram colectados mais de 21 milhões de recursos e quase 600 milhões de ligações entre páginas. Este problema está a ser estudado e estão a ser ensaiadas estratégias de distribuição e partição da base de dados.

4.2 Exploração de Dados

O RDC é baseado em acrescentos e modificações realizados sobre o Larbin, concretizado em C++ que foi usado como ponto de partida. Para a obtenção da informação desejada, não disponível na versão original, foram feitas um conjunto de modificações bem definidas e localizadas, no sentido de possibilitarem a utilização de novas versões do Larbin, que está disponível em domínio público. A informação a registar é bastante mais ampla que a utilizada neste trabalho, e foi definida no âmbito do Projecto NetCensus [2].

Uma primeira constatação é o acesso à informação do DNS. Para cada *host* visitado, pretende-se obter os seus vários nomes alternativos e endereços IP em que está acessível. Para esse feito foi necessário registar no módulo do aDNS [7] a informação obtida do serviço de resolução de nomes. Uma segunda modificação necessária foi o registo dos tempos de descarga de conteúdos. Esses tempos são imprescindíveis no âmbito do NetCensus, e não são disponibilizados pelo Larbin. Outros problemas que se teve que ultrapassar foi a manipulação dos URLs com *redirects*, devolvidos como situação de erro pelo Larbin, e o processamento criterioso das situações de erro. O procedimento do utilizador invocado pelo Larbin foi concretizado no sentido da obtenção das estatísticas de todos conteúdos e a gravação em disco local de um conjunto, definido em tempo de configuração, de tipos e subtipos MIME. Para cada conteúdo é gravado um registo XML com toda a informação para a BDP. Foi também necessário sincronizar a gravação de ficheiros de dados XML com a gravação do estado do robot para recuperação em caso de falha.

O Larbin vai ser modificado para suportar uma versão distribuída, em que vários robots cooperam e colectam partições do espaço de informação. Cada robot vai usar, para além disso, uma tabela de *hash* distribuída suportada por um *cluster* de PCs. Desta forma há dois níveis de distribuição: ao nível da rede local e na rede de longa distância.

O GTrace foi modificado para permitir a localização de lotes de *hosts* em alternativa à interface gráfica usada para a resposta a pedidos individuais. Na nova versão, dado um ficheiro com nomes de hosts e respectivos endereços IP, é devolvido um ficheiro que contém um registo por *host*, com o nome, endereço IP, latitude, longitude e localização geográfica. Um problema que houve que resolver foi a delicadeza da cadência de pedidos realizados ao servidor NetGeo, mantido pela CAIDA [8]. Nesse sentido optou-se por pedidos com vários nomes de *hosts*. O GTrace está concretizado em Java e é disponibilizado pela CAIDA.

Para análise de conteúdos para extracção de informação de localização, num primeiro passo os conteúdos vão ser etiquetados usando um analisador morfológico. Só um subconjunto de palavras como nomes ou números de telefone serão confrontados com os diversos tipos de localização geográfica.

4.2 A Visualização de Resultados

Foi usado como ponto de partida o SIG do *GeoMedia Web Map* e desenvolvidas um conjunto de VBScripts (*Visual Basic Scripts*) para a apresentação de resultados. Utilizando tecnologia ASP (*Active Server Pages*) foi possível integrar código HTML com as *scripts* necessárias ao processamento dos mapas temáticos desejados.

Os módulos desenvolvidos permitem integrar a informação processada pelo sistema GeoCensus com a informação cartográfica que integra o SIG. Desta forma, como resposta aos diversos tipos de interrogação, é devolvida uma tabela com um ou mais registos, os quais são devidamente manipulados por forma a serem visualizados graficamente.

Neste fase da implementação do sistema e uma vez que apenas se encontra finalizada a localização de servidores, a visualização de resultados está limitada à distribuição geográfica de *hosts*. É possível verificar a distribuição dos servidores pelo país, para um dado distrito, ou ainda, visualizar a localização de um servidor específico dado o seu IP. A Figura 9 apresenta a interface desenvolvida para a apresentação de resultados. Na mesma o utilizador selecciona a opção pretendida e indica, se for o caso, o distrito ou o IP a localizar.

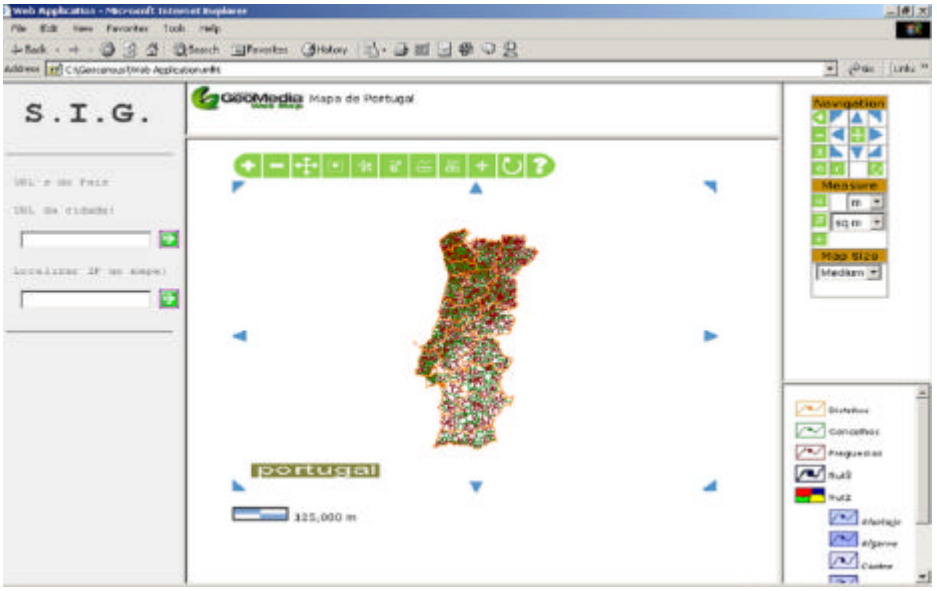

Figura 9 – Visualização de Resultados

A título de exemplo, a Figura 10 apresenta a distribuição quantitativa de servidores para o distrito do Porto. Chama-se a atenção para o facto dos testes à interface gráfica terem sido efectuados com dados fictícios, pelo que a distribuição apresentada não corresponde à realidade.

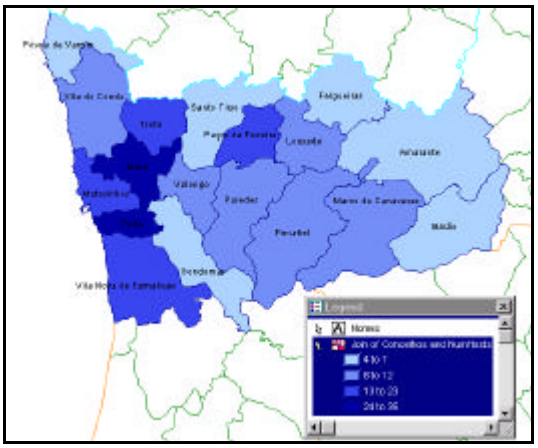

Figura 10 – Distribuição de servidores no distrito do Porto

5 APRESENTAÇÃO DE RESULTADOS

Cidade	Hosts	Ips	Páginas	Ref.Ent.	Ref. Sai.
COIMBRA	75	77	39977	922238	668550
LISBOA	102	102	22040	489449	458480
BRAGA	50	50	14580	226553	215679
PORTO	32	32	11970	334711	325150
BEJA	8	8	3900	150654	137177
AVEIRO	9	9	2722	100557	97406
BRAGANCA	7	7	1945	50521	48429
LEIRIA	8	8	1356	54174	45056
CAMPO GRANDE	25	25	1331	20637	15609
VILA REAL	9	9	1094	15795	13336
VIANA DO CASTELO	7	7	617	11550	10816
VISEU	7	7	515	12391	10756
OEIRAS	2	2	476	4998	4857
ALGES	18	18	359	6143	5118
CASTELO BRANCO	9	9	196	4105	2662
VILA NOVA DE GAIA	1	1	179	11257	10762
PORTIMAO	1	1	168	3274	3115
GUARDA	3	4	142	2407	2077
SANTAREM	1	1	43	432	414
MAFRA	1	1	36	259	227
TOMAR	3	3	36	412	389
CANTANHEDE	2	2	34	327	279
MAIA	1	1	30	477	438
TAVIRA	1	1	27	184	132
FARO	3	3	19	747	709
CALDAS DA RAINHA	1	1	18	728	674
SETUBAL	1	1	16	242	234
TORRES VEDRAS	1	1	12	365	327
GUIMARAES	3	3	6	59	100
POVOA DE VARZIM	1	1	3	19	12
CAPARICA	1	1	3	15	11
PORTALEGRE	2	2	3	143	143
BARCELOS	1	1	1	3	4
Total	396	399	103854	2425826	2079128

Tabela 1- Algumas estatísticas da amostra analisada

O ênfase deste artigo é a descrição da arquitectura do sistema GeoCensus, a apresentação das funcionalidades disponíveis nos seus diferentes componentes e a respectiva interacção. Os dados apresentados são apenas ilustrativos e a sua apresentação tem como objectivo mostrar que o sistema tem as potencialidades necessárias para os fins a que se destina. A apresentação dos dados relativos a uma amostra significativa do Web em Portugal está reservada para trabalhos futuros.

Foi objecto de análise um subconjunto duma amostra da Web em Portugal cuja colecta foi realizada no início de 2002. Da amostra foram seleccionados 365 servidores Web pertencentes a estabelecimentos de ensino superior e Câmaras Municipais. Foi feita esta escolha para facilitar a existência duma referência de base para a localização geográfica das páginas.

Neste trabalho apenas se explorou a localização geográfica dos computadores hospedeiros das páginas e a sua relação na topologia Web. A extracção de identificadores geográficos do conteúdo das páginas foi deixado para trabalho futuro.

As estatísticas da pequena amostra que foi objecto de análise são apresentadas na Tabela 1. Os dados estão ordenados pelo número de páginas. Os dados correspondentes às referências de saída apontam para recursos que podem estar fora do espaço em análise. Ao contrário, as referências de entrada são apenas para páginas que integram a amostra analisada.

Como não podia deixar de ser, a ordem corresponde às principais cidades portuguesas. Existe uma aparente uma relação de proporcionalidade entre o número de servidores, páginas e referências de entrada e de saída.

Na Tabela 2 apresenta-se a informação de localização geográfica de um conjunto de páginas seleccionadas. Embora esses dados não sejam representativos, contêm a informação necessária para o cálculo do âmbito geográfico estimado das páginas referenciadas, se forem combinados com a informação apresentada na Tabela 1, que nos dá o total de páginas por localização.

Url	http://www.up.pt:80/	http://www.ua.pt:80/	http://www.fe.up.pt:80/	http://www.utad.pt:80/	http://www.uc.pt:80/	http://www.uevora.pt:80	http://www.ubi.pt:80/	http://www.ul.pt:80/
VILA REAL				83	1			
AVEIRO	2	1136	2		9	2		
PORTO	213	14	353	3	9	5	1	1
VIANA DO CASTELO				3				
COIMBRA	1	2	3	2	166	4	4	2
BRAGA	5	3	12	1	2	1	1	1
ALLENTOWN					1			1
OEIRAS					1			15
BALTIMORE							3	
MAIA	3	4	1	3	10	5	3	3
LISBOA	1	1	1	1	1	1	1	1
CASTELO BRANCO					6			
VISEU	2	2	1	13	2	1	2	2

TABELA 2: Localização geográfica das referências que apontam para um conjunto de páginas

Como se usou uma amostra de pequena dimensão, sentido calcular o âmbito geográfico destas páginas.

De qualquer forma isto demonstra que é possível estimar o âmbito geográfico para cada página do espaço Web, usando apenas a localização geográfica dos hosts que as hospedam e a sua relação na topologia Web.

6 CONCLUSÕES

Neste artigo foi apresentado o sistema GeoCensus, o qual aborda a extracção de informação geográfica na Web. Apresentada a sua arquitectura e as principais funcionalidades do sistema, é descrito com particular detalhe o processo de localização geográfica de *hosts*. Com base nesta localização pode-se calcular o AGE para as páginas.

A qualidade da informação geográfica pode ser melhorada se combinar a informação extraída desta forma com outra fonte de informação que é conteúdo das próprias páginas.

7 AGRADECIMENTOS

A realização deste trabalho contou com a participação da Natália Vivas e da Nelma Mogas da equipa de trabalho do projecto GeoCensus, e ainda com a colaboração do Leopoldo Silva, membro da equipa de trabalho do projecto NetCensus. O projecto GeoCensus está a ser subsidiado pelo Centro de Investigação Algoritmi.

8 REFERÊNCIAS

1. ISO/TC-211, Geographic Information - Spatial referencing by geographic identifiers. Document ISO/TC 211 N 1172, 2001, International Standard Organisation.

2. Silva, L.O., et al. NetCensus: Medição da evolução dos conteúdos na Web. in Actas da 5a. Conferência sobre Redes de Computadores - Protocolos, Tecnologias e Aplicações Rumo à Internet. 2002. Faro, Portugal.

3. Periakaruppan, R. and E. Nemeth. GTrace - A Graphical Traceroute Tool. in 13th. Systems Administration Conference, LISA'99. 1999. Seattle, USA.

4. Ding, J., L. Gravano, and N. Shivakumar. Computing Geographical Scopes of Web Resources. in proceedings of the 26th. VLBD Conference. 2000. Cairo, Egypt.

5. Intergraph, GeoMedia Web Map v4.0 Online Documentation, . 2001, Intergraph Corporation.

6. Larbin, Larbin Web Crawler, http://www.gnu.org/directory/network/tools/larbin.html. 2002.

7. aDNS, G., Advanced, easy to use, asynchronous-capable DNS client library and utilities, http://www.chiark.greenend.org.uk/~ian/adns. 2002.

8. Caida, NetGeo - The Internet Geographic Database, http://www.caida.org/tools/utilities/netgeo . 2002.

Joaquim MACEDO
macedo@di.uminho.pt

Joaquim Macedo é Professor Auxiliar do Departamento de Informática da Escola de Engenharia da Universidade do Minho. Licenciado em Engenharia Electrotécnica na Universidade Agostinho Neto, Angola, em 1983, obteve o seu doutoramento em Informática na Universidade do Minho em 2002.

Universidade do Minho
Departamento de Informática
Campus de Gualtar
4710 Braga
Tel: (+ 351) 253 604470
Fax: (+ 351) 253 604471
Email: macedo@di.uminho.pt
URL: http://www.di.uminho.pt/~macedo

Maribel Yasmina SANTOS
maribel@dsi.uminho.pt

Maribel Yasmina Santos é Professora Auxiliar do Departamento de Sistemas de Informação da Escola de Engenharia da Universidade do Minho. Licenciada em Engenharia de Sistemas e Informática, pela Universidade do Minho em 1991, obteve o grau de mestre em Informática (especialização em Informática de Gestão) pela mesma Universidade em 1996. Em 2001 concluiu, na Universidade do Minho, o doutoramento em Tecnologias e Sistemas de Informação, área de conhecimento de Engenharia da Programação e dos Sistemas Informáticos.

Universidade do Minho
Departamento de Sistemas de Informação
Campus de Azurém
4800-058 Guimarães

Tel: (+ 351) 253 510259
Fax: (+ 351) 253 510250
Email: maribel@dsi.uminho.pt
URL: http://www.dsi.uminho.pt/~maribel