konferenca COBISS 2014

PRVA ITERACIJA ZAJEMA SLOVENSKE DOMENE .SI

izzivi, pasti in ovire

Matjaž Kragelj, Narodna in univerzitetna knjižnica

NARODNA IN UNIVERZITETNA KNJIŽNICA

OBVEZNI IZVOD SPLETNIH PUBLIKACIJ

- Spletne publikacije so
  „elektronske knjige, elektronski časopisi in časniki, dostopni po spletu ter spletne strani in podobno“
  Zakon o obveznem izvodu publikacij (Ur. list RS št. 69/06 – ZOIPub)

- **Pravilnik** o vrstah in izboru elektronskih publikacij za obvezni izvod med spletne publikacije uvršča tudi:

  - spletna mesta in portale organizacij, oseb in dogodkov,
  - spletno dostopne storitve, podatkovne zbirke, novice,
  - forume, spletne biltene (newsletters)
  - druge elektronske vsebine kot so video in zvočni zapisi,
  - interaktivni zemljevidi in mestni načrti, itd
  Pravilnik (Ur. list RS št. 90/07)

SELEKTIVNI ZAJEM SPLETNIH MEST

November 2014 (arhiv.nuk.uni-lj.si)

- 1230+ spletnih domen različnih področij
- 10 TB podatkov
- 70.196.529 indeksiranih spletnih strani
  - iskalne po polnem besedilu
  - SOLR index
  - fasete / filtri
  - Iskanje podobnih (SOLR MoreLikeThis)

ZAJEM SPLETNIH DOMEN *.si*

DEJSTVA:

- Obstoječih domen ***.si*** ≈ 105.000 (vir: Arnes)
- Aktivnih domen ≈ 75%
- Neobstoječih domen ≈ 10%
- Neodzivnost strežnika ali blokada ICMP (*Internet Control Message Protocol)* ≈ 15%

PRIČETEK ZAJEMANJA:

- april 2014
  Prednost zajemanja - domene *„pred potekom“*

ZAJEM SPLETNIH DOMEN *.si*

TRENUTNO STANJE (november 2014):

- zajetih ≈ 40% vseh domen
- zajetih dokumentov: 25.296.414
- količina ≈ 1.3 TB

STRATEGIJA ZAJEMA:

Prednost imajo domene, ki bodo kmalu potekle

- zajem 3 nivoje globoko
- omejitev zajema: 500mb / domeno
- omejitev zajema: 10.000 dokumentov

ZAJEM SPLETNIH DOMEN *.si*

DESET „NAJTEŽJIH MIME-TIPOV“ (v bajtih)

image/jpeg	418.911.127.930
text/html	372.742.904.570
application/pdf	150.185.303.650
video/mp4	120.153.188.037
image/png	52.582.408.396
audio/mpeg	17.865.616.919
video/x-flv	15.947.500.439
application/x-shockwave-flash	*12.774.687.430*
text/plain	8.992.526.528
application/x-javascript	*8.606.425.340*

ZAJEM SPLETNIH DOMEN *.si*

DESET „NAJ pojavitev MIME-TIPOV“

text/html

image/jpeg

image/png

image/gif

text/css

application/javascript

text/dns

text/xml

application/x-javascript

text/plain

DESET PRIMEROV NEVELJAVNIH MIME TIPOV

<?=image/pjpeg
null
image/$type
#<Mime::NullType:0xbeee6fc4>
/
1
ďż˝ďż˝ďż˝application/x-woffďż˝ďż˝ďż˝
<?=image/gif
System.Byte[]
Peter

ZAJEM SPLETNIH DOMEN *.si*

KONCEPTUALNE OVIRE

DOSTOP DO PODDOMEN

- dostop možen preko administratorja domene (pridobitev seznam poddomen)
- iskanje nizov poddomene v zajemanju domene (hevristika)

Primer:

gov.si, uni-lj.si, *.si

arso.gov.si, *ess*.gov.si, *prostor*.gov.si, *gu*.gov.si, *mddsz*.gov.si, *nuk*.uni-lj.si…

ZAJEM SPLETNIH DOMEN *.si*

TEHNIČNE OVIRE

- Generacija vsebine s pomočjo skriptnih jezikov
  - AJAX tehnike
    (Forum komentarji, Twitter, Gmail, itd.)
  - Flash plugin, Flash loader, Java script
    - pajek *ne zna interpretirati vsebine*
    - *ne zna izvesti funkcije*, ki dopolni strani
      (asinhroni prenosi)
    - *ne zna uporabiti rtsp* protokola
      (real time streaming protocol)

ZAJEM SPLETNIH DOMEN *.si*

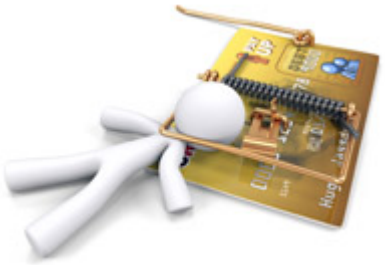
TEHNIČNE OVIRE

Nudenje video vsebin:

- Video kot *objekt*, *datoteka*
  (prenos ali predvajanje z brskalnikom s pomočjo *HTML5* funkcij ali plugin-ov
- Video kot vir (stream)
  (pajek / robot ne zna dostopati do vira vsebine)

ZAJEM SPLETNIH DOMEN *.si*

PASTI



- pasti za pajke
  - dinamično generiranje novih URL naslovov (absolutne iz relativnih poti, rezultat: *novi url-ji* v neskončnost)
  - koledarji, trgovina - sortiranje, razvrščanje (rezultat: *novi url-ji*)
  - dodeljevanje seje zaradi neuporabe piškotkov (session variable, rezultat: *novi url-ji*)

ZAJEM SPLETNIH DOMEN *.si*

REŠITVE (tips & tricks)

- Implementacija pajka, ki simulira delovanje brskalnika (npr. *PhantomJS*)
  - Primer: *socialna omrežja* – Twitter. sestavljanje vsebine v celoto – „append“ (page1+page2+..+page n)

  Podobno lahko rešujemo forume, gmail, itd..
- *Integracija rtsp* protokola v pajka / robota za zajem.
- *Razvoj dodatne logike* za zajem (video) vsebin iz spletnih mest

Hvala za pozornost!

matjaz.kragelj@nuk.uni-lj.si