

Where is the user? Filtering Bots from the Edurep Query Logs

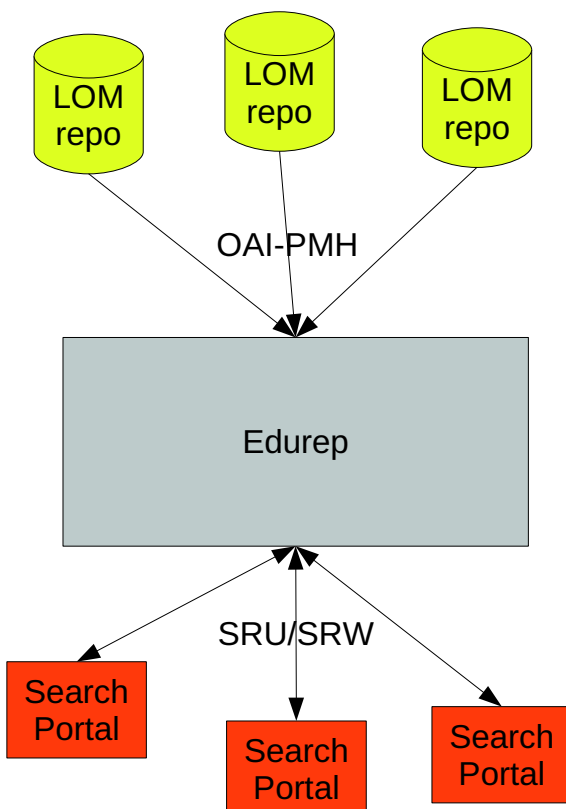


Wim Muskee

Kennisnet Foundation

SE@M 2010

Edurep in Context



- harvesting more than 50 repositories
- almost 800.000 records
- queried by more than 10 production portals
- more than 500.000 queries each month (peak > 2.000.000)

Query Language

- CQL 1.1 over SRU/SRW
- query=elephant
- query=elephant OR mammoth
- query=lom.general.title=elephant
- query=snake AND lom.educational.context=PO
- query=lom.lifecycle.contribute.publisherdate>2009

more query examples

- Pagination:
query=dog
query=dog&startRecord=11&maximumRecords=10
query=dog&startRecord=21&maximumRecords=10
- Specific records:
query=lom.general.catalogentry=memorix:na:col1:dat148993
query=meta.upload.id=na:oai:mdms.kenict.org:217136
- Facets using facet index:
query=dog&x-term-drilldown=lom.technical.format

XML Query Response

```
- <searchRetrieveResponse>
  <swversion>1.1</swversion>
  <swnumberOfRecords>20</swnumberOfRecords>
- <swrecords>
- <swirecord>
  <swirecordSchema>100</swirecordSchema>
  <swirecordPackaging>xml</swirecordPackaging>
- <swirecordData>
- <cp:location schemaLocation="http://www.imagingtool.org/schemas/zip/1/mond_v1/p1.xsd">
- <cp:general>
- <cp:title>
  <cp:langstring xml:lang="nl">Fossiele slang zo lang als T-rex</cp:langstring>
</cp:title>
- <cp:catalogentry>
  <cp:catalog>tag</cp:catalog>
- <cp:entry>
  <cp:langstring xml:lang="en">tag:www.kennislink.nl:2009:Publication:3336</cp:langstring>
</cp:entry>
</cp:catalogentry>
<cp:language>nl</cp:language>
- <cp:description>
- <cp:langstring xml:lang="nl">
  In Colombia is de grootste slang ooit gevonden. Het fossiel is naar schatting dertien meter lang. De slang is als oerthermometer gebruikt voor de tropen van zo'n 60 miljoen
  jaar geleden. Het was toen ruim warmer dan tegenwoordig. Worden de kwetsbare tropen nog warmer als de aarde verder opwarmt?
</cp:langstring>
</cp:description>
- <cp:keyword>
  <cp:langstring xml:lang="nl">Nieuws</cp:langstring>
</cp:keyword>
- <cp:keyword>
  <cp:langstring xml:lang="nl">waarde & kijkmoment</cp:langstring>
</cp:keyword>
- <cp:keyword>
```

Search Portals

Wikiwijs Zoek, maak & deel leer materiaal

home pe sbo vo so mbo ho maken delen

Zoek op trefwoord

Wikiwijs home - Zoeken: slang

Huidige zoekopdracht: slang

Trefwoorden: slang

Verfijn 206 resultaten

Er zijn 206 resultaten gevonden

1 Fossiele slang zo lang als T-rex

Gratis ja Leefbaarheidsgraad: 13

In Colombia is de grootste slang ooit gevonden. Het fossiel is naar schatting dertien meter lang. De slang is als oerthermometer gebruikt voor de tropen van zo'n 60 miljoen jaar geleden. Het was toen ruim warmer dan tegenwoordig. Worden de kwetsbare tropen nog warmer als de aarde verder opwarmt?

2 Statische elektriciteit (expr_id: 231) : Een stroomslang

Gratis ja Leefbaarheidsgraad: 13

Dit wordt een hele lange slang. Die slang leg je op een aluminiumplaat. Door de pen langs de staal te wrijven komt er een beetje stroom in de pen. En door die stroom komt de slang ontloosd. Zo zijn slangen ontworpen om...

3 Slangen : Hoe ziet een slang eruit?

Gratis ja Leefbaarheidsgraad: 7

Slangen zijn reptielen zonder poten. De huid is bedekt met schubben en ze leggen eieren.

EduRep CONTENTtoomer

slang

Resultaten 1-5 van 206 voor slang

Fossiele slang zo lang als T-rex

In Colombia is de grootste slang ooit gevonden. Het fossiel is naar schatting dertien meter lang. De slang is als oerthermometer gebruikt voor de tropen van zo'n 60 miljoen jaar geleden. Het was...

Statische elektriciteit (expr_id: 231) : Een stroomslang

Dit wordt een hele lange slang. Die slang leg je op een aluminiumplaat. Door de pen langs de staal te wrijven komt er een beetje stroom in de pen. En door die stroom komt de slang ontloosd. Nog een...

Slangen : Hoe ziet een slang eruit?

Slangen zijn reptielen zonder poten. De huid is bedekt met schubben en ze leggen eieren.

Engel dieren kijken : Met Flip de beer

Kinderen en Flip bekijken slangen en spinnen achter glas. Maar ze durven ook een grote spin in hun hand vast te houden en een slang om hun nek te hebben.

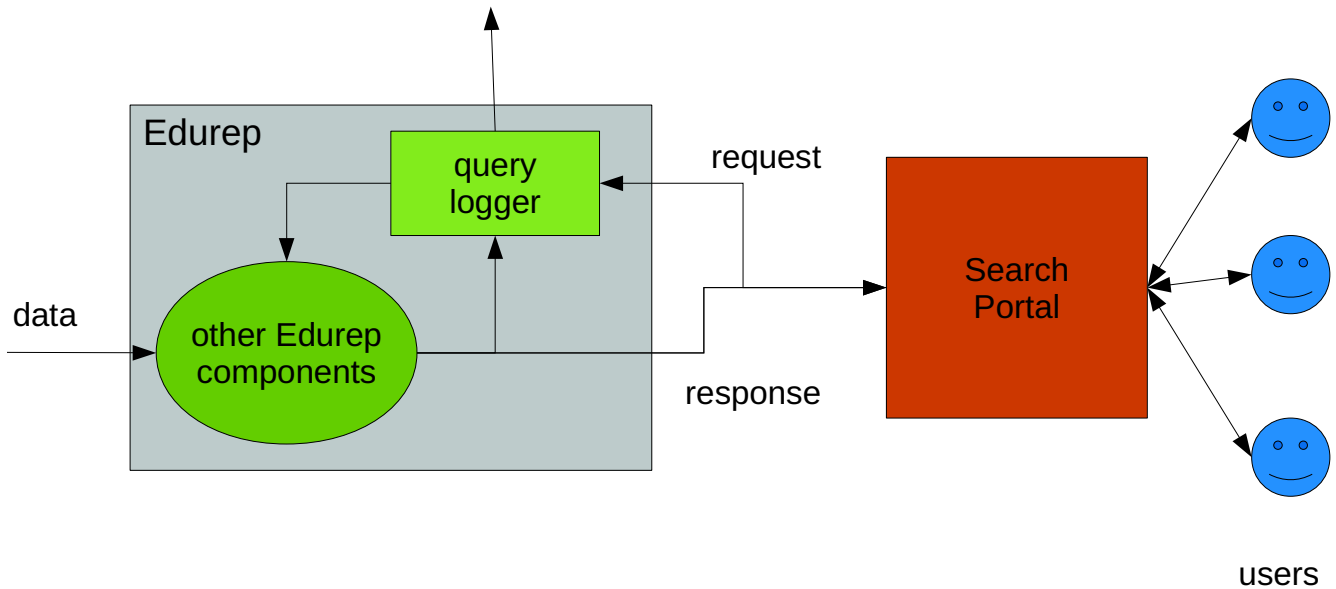
Voeren van een slang met een levend prooidier

Werkwijzer over het voeren van een slang met een levend prooidier. Aandachtspunten zijn de activiteit van de slang en de controle.

[1 | 2 | 3 | 4 | 5 | 6 | 7 | 8]

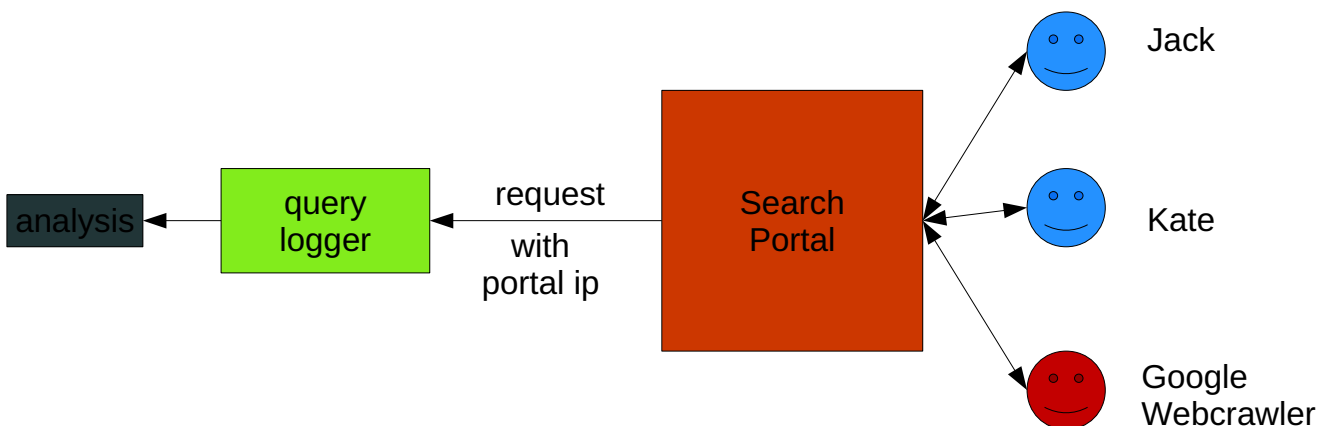
Query Logging

<datetime> <ip> <response size> <response time> <protocol> <query>



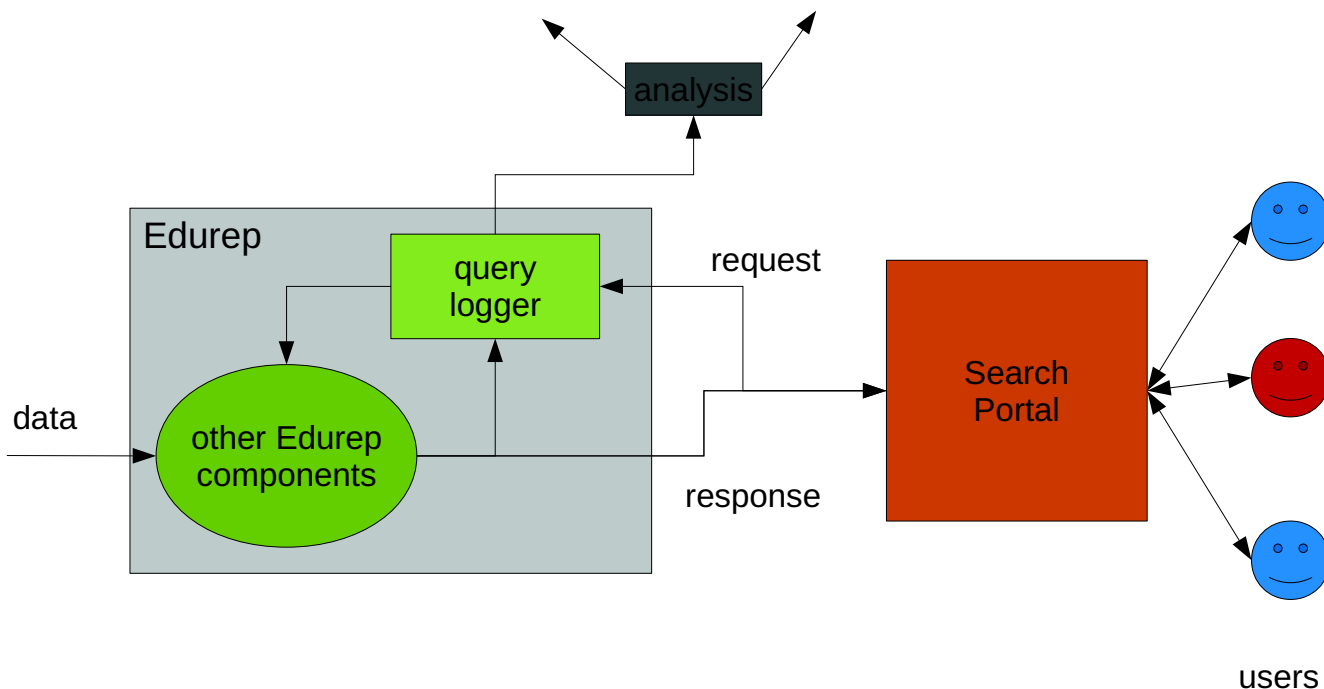
Problem

- Human and automated users on search portals.
 - extra traffic
- At Edurep level, regular means of user type distinction cannot be used.
 - no reliable user behaviour



Goal

- Distinguish **automated** and **user** generated queries



Hyperlinks

- Webcrawlers automatically follow hyperlinks.
- 4 types:
 - **search link**: retrieve a first page resultset
 - **pagination link**: retrieve another resultset page
 - **result link**: retrieve a specific record
 - **facet link**: retrieve the nr of records for that facet
- Each link type has a specific query pattern.

Search Link

retrieve a first page resultset

Totaal aantal gevonden leermiddelen: 13899

- Fit-it**
Materiaal voor het maken van woordjes. Het materiaal bestaat uit vijftien gele houten blokjes met zw...
- Mijn**
Lekserie bij Veilig Leren Lezen, 1e en 2e maandvak.
- Lustertaal**
Materiaal dat bestaat uit: een opzetboek met zestien gebouwen en een gleuf, acht opzetplaten, pionnen, ...
- Spellinggedrag en spellingproblemen**
Materiaal waarmee leerkrachten zicht kunnen krijgen op de denkprocessen die ten grondslag liggen aan...
- Groepboek voor Windows**
Hulpprogramma voor o.a. "Werkwoordenboek voor Windows" geheel computergebruikt in de klas te gebruik...
- Op niveau tweede fase**
Methode Nederlands voor de tweede fase. De methode bestaat uit één havo/vwo informatieboek en apar...
- ABC-woordspel : spel voor technisch lezen**
Vier spelen voor het bevorderen van het voortgezet technisch lezen. Het materiaal bestaat uit: vier...
- Interventiepakket dyslexie praktijkonderwijs**
Interventieprogramma voor het begeleiden en coachen van leerlingen met handrekkige lees- en spelling...
- Zestien plus**
Een doorstromingsgericht basis- en beginnersprogramma Nederlands als tweede taal voor volwassenen. D...
- Start-krant : de belangrijkste krant van Nederland**
De Start-krant is een krant voor mensen die moeite hebben met (Nederlands) lezen. Hiertoe behoren o...

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 volgende pagina >

Filter op label:

Soort Leermateriaal

informatiebron (9008)
gekleed opdracht (2251)
open opdracht (1743)
handreiking (1300)
evaluatie- en toetsmateriaal (341)
praktisch (191)
vergelijken (38)
digitaalverhaal (36)
verkenning- en onderzoeksmateriaal (30)
lezen (26)

Besogde eindgebruiker

leerder (8598)
leerkracht (3771)

Onderwijstype

primaire onderwijs (9072)
voortgezet/secondair onderwijs (3424)
voor- en vroegschoolse educatie (2248)
speciaal/buiten gewoon basisonderwijs (1132)
volwassenenonderwijs (895)
speciaal onderwijs (292)

Search Link

retrieve a first page resultset

- startRecord == 1
- OR
- startRecord == null (using default 1)
- AND
- maximumRecords >= 5
- OR
- maximumRecords == null (using default 10)

Pagination Links

retrieve another resultset page

Totaal aantal gevonden leermiddelen: 13899

- Fit-it**
Materiaal voor het maken van woordjes. Het materiaal bestaat uit vijftien gele houten blokjes met zw...
- Haan**
Leesserie bij Vrijlg Leren Lezen, 1e en 2e maandversie.
- Lustertaal**
Materiaal dat bestaat uit: een opzetblok met zestien gaas en een gleuf, acht opzetplaten, pionnen, ...
- Spelviggedrag en spelingsproblemen**
Materiaal waarmee leerkrachten zich kunnen krijgen op de denkprocessen die ten grondslag liggen aan...
- Groepsboek voor Windows**
Hulpprogramma voor o.a. "Woordboek voor Windows" geheel computergestuurd in de klas te gebruik...
- Op niveau tweede fase**
Methode Nederlands voor de tweede fase. De methode bestaat uit één havo/vwo informatieboek en apar...
- ABC-woordspel : spel voor technisch lezen**
Vier spelen voor het bevorderen van het voortgezet technisch lezen. Het materiaal bestaat uit: vier...
- Interventiepakket dyslexie praktijkonderwijs**
Interventieprogramma voor het begeleiden en coachen van leerlingen met handrekkige lees- en spelling...
- Zestien plus**
Een doorstromingsgericht basis- en beginnersprogramma Nederlands als tweede taal voor volwassenen. D...
- Start-kraak : de begripelijkste kraak van Nederland**
De Start-kraak is een kraak voor mensen die moeite hebben met (Nederlands) lezen. Hiertoe behoren o...

3 | 2 | 7 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 volgende pagina »

zoek

Filter op label:

Soort Leermateriaal

informatiebron (9008)
gekleefd opdracht (2251)
open opdracht (1743)
handreiking (1308)
evaluatie- en toetsmateriaal (341)
gemeenschap (191)
vergelijken (38)
digitaalverhaal (35)
verkenings- en onderzoeksmateriaal (30)
lezen (28)

Beoogde eindgebruiker

leider (8599)
leerkracht (3771)

Onderwijstype

primaire onderwijs (9072)
voortgezet/secondair onderwijs (3424)
voor- en vroegschoolse educatie (2248)
speciaal/buitengewoon basisonderwijs (1132)
volwassenenonderwijs (895)
speciaal onderwijs (292)

Pagination Links

retrieve another resultset page

- startRecord > 1

Result Links

retrieve a specific record

Totaal aantal gevonden leermiddelen: 13899

1. **Fit-it**
Materiaal voor het maken van woordjes. Het materiaal bestaat uit vijftien gele houten blokjes met zw...
2. **Helen**
Leesserie bij Vrijg Lezen Lezen, 1e en 2e maandserie.
3. **Lustertaal**
Materiaal dat bestaat uit: een opzetblok met zestien gaas en een gleuf, acht opzetplaten, pionnen, ...
4. **Spelviggedrag en spelingsproblemen**
Materiaal waarmee leerkrachten zich kunnen krijgen op de denkprocessen die ten grondslag liggen aan...
5. **Groepsboek voor Windows**
Hulpprogramma voor o.a. "Woordboek voor Windows" geheel computerbestuurd in de klas te gebruik...
6. **Op niveau tweede fase**
Methode Nederlands voor de tweede fase. De methode bestaat uit één havo/vwo informatieboek en apar...
7. **ABC-woordspel : spel voor technisch lezen**
Vier spelen voor het bevorderen van het voortgezet technisch lezen. Het materiaal bestaat uit: vier...
8. **Interactiepakket dyslexie praktijkonderwijs**
Interventieprogramma voor het begeleiden en coachen van leerlingen met hardnekkige lees- en spelling...
9. **Zestien plus**
Een doorstromingsgericht basis- en beginnersprogramma Nederlands als tweede taal voor volwassenen. D...
10. **Start-krant : de belangrijkste krant van Nederland**
De Start-krant is een krant voor mensen die moeite hebben met (Nederlands) lezen. Hiertoe behoren o...

3 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 volgende pagina »

Filter op label:

Soort Leermateriaal

informatiebron (9008)
gekleefd opzacht (2251)
open opdracht (1743)
handleiding (1308)
evaluatie- en toetsmateriaal (341)
gemeenschap (191)
vergelijken (38)
digitaalverhaal (36)
verkenings- en onderzoeksmateriaal (30)
lezen (28)

Beoogde eindgebruiker

leider (8599)
leerkracht (3771)

Onderwijstype

primaire onderwijs (9072)
voortgezet/secondair onderwijs (3424)
voor- en vroegschoolse educatie (2248)
speciaal/buitengewoon basisonderwijs (1132)
volwassenenonderwijs (895)
speciaal onderwijs (292)

Result Links

retrieve a specific record

- startRecord == 1
- OR
- startRecord == null (using default 1)
- AND
- query on unique record aspect (eg. catalogentry)

Facet Links

retrieve the amount of records for that facet

Totaal aantal gevonden leermiddelen: 13899

- Fit-it**
Materiaal voor het maken van woordjes. Het materiaal bestaat uit vijftien gele houten blokjes met zw...
- Hazen**
Leesserie bij Vrijig Leren Lezen, 1e en 2e maandversie.
- Lustertaal**
Materiaal dat bestaat uit: een opzetboek met zestien gaas en een gleuf, acht opzetplaten, pionnen, ...
- Spelviggedrag en spelingsproblemen**
Materiaal waarmee leerkrachten zich kunnen krijgen op de denkprocessen die ten grondslag liggen aan...
- Groepsboek voor Windows**
Hulpprogramma voor o.a. "Woordboek voor Windows" geheel computergestuurd in de klas te gebruik...
- Op niveau tweede fase**
Methode Nederlands voor de tweede fase. De methode bestaat uit één havo/vwo informatieboek en apar...
- ABC-woordspel : spel voor technisch lezen**
Vier spelen voor het bevorderen van het voortgezet technisch lezen. Het materiaal bestaat uit: vier...
- Interventiepakket dyslexie praktijkonderwijs**
Interventieprogramma voor het begeleiden en coachen van leerlingen met handrekkige lees- en spelling...
- Zestien plus**
Een doorstromingsgericht basis- en beginnersprogramma Nederlands als tweede taal voor volwassenen. D...
- Start!-krant : de belangrijkste krant van Nederland**
De Start!-krant is een krant voor mensen die moeite hebben met (Nederlands) lezen. Hiertoe behoren o...

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 volgende pagina »

Filter op label:

Soort Leermateriaal

informatiebron (9008)
gekluisd opdracht (2251)
open opdracht (1743)
handreiking (1308)
evaluatie- en toetsmateriaal (341)
gemeenschap (191)
vergelijken (38)
digitaalverhaal (36)
verkennings- en onderzoeksmateriaal (30)
leem (26)

Beoogde eindgebruiker

leider (8599)
leerkracht (3771)

Onderwijstype

primaire onderwijs (9072)
voortgezet/secondair onderwijs (3424)
voor- en vroegschoolse educatie (2248)
speciaal/buitengewoon basisonderwijs (1132)
volwassenenonderwijs (895)
speciaal onderwijs (292)

Facet Links

retrieve the amount of records for that facet
(using facet index)

- query
- AND
- x-term-drilldown

```
-<srw:extraResponseData>
- <dd:drilldown xsi:schemaLocation="http://ineresco.org/namespace/drilldown http://m
- <dd:term-drilldown>
- <dd:navigator name="lom.educational.context.value">
  <dd:item count="769">VO</dd:item>
  <dd:item count="241">FO</dd:item>
  <dd:item count="181">BVE</dd:item>
  <dd:item count="30">VVE</dd:item>
  <dd:item count="26">primaire onderwijs</dd:item>
  <dd:item count="21">voortgezet onderwijs</dd:item>
  <dd:item count="12">beroepsonderwijs en volwasseneneducatie</dd:item>
  <dd:item count="9">speciaal onderwijs</dd:item>
  <dd:item count="9">speciaal basisonderwijs</dd:item>
  <dd:item count="7">other</dd:item>
</dd:navigator>
- <dd:navigator name="lom.rights.cost">
  <dd:item count="951">no</dd:item>
  <dd:item count="58">yes</dd:item>
</dd:navigator>
- <dd:navigator name="lom.general.aggregationLevel">
  <dd:item count="931">1</dd:item>
  <dd:item count="153">2</dd:item>
  <dd:item count="78">4</dd:item>
  <dd:item count="38">3</dd:item>
</dd:navigator>
- <dd:navigator name="lom.technical.format">
  <dd:item count="741">text/html</dd:item>
  <dd:item count="584">image/jpeg</dd:item>
  <dd:item count="318">video/x-ms-asf</dd:item>
  <dd:item count="31">application/zip</dd:item>
  <dd:item count="24">application/pdf</dd:item>
  <dd:item count="16">non-digital</dd:item>
  <dd:item count="5">application/msword</dd:item>
  <dd:item count="4">text/xml</dd:item>
  <dd:item count="2">application/octet-stream</dd:item>
  <dd:item count="2">image/gif</dd:item>
</dd:navigator>
</dd:term-drilldown>
</dd:drilldown>
</srw:extraResponseData>
```

Facet Links

retrieve the amount of records for that facet
(not using facet index)

- startRecord == 1 OR null (using default 1)
AND
- maximumRecords == 0 OR 1*
AND
- ! query on unique record aspect (eg. catalogentry)
foreach
- query=lom.field=value
(to retrieve the nr of records for that value)

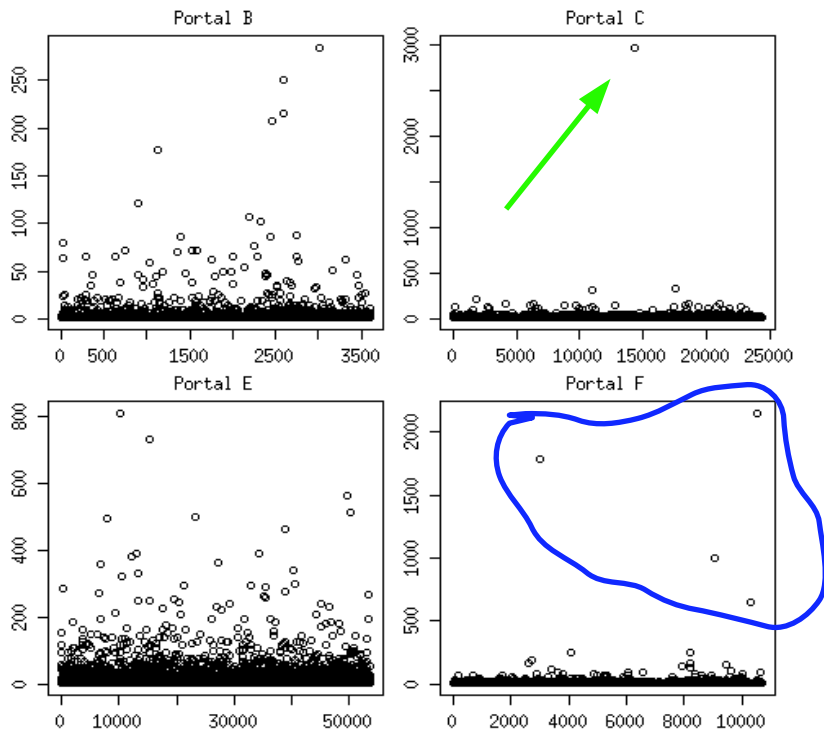
Dataset

- Januari 2010 logs
- 6 largest portals
- 5 variables from each query
 - portal ip
 - startRecord
 - maximumrecords
 - search query hash
 - unique request boolean

portal	requests
A	41.690
B	126.340
C	1.293.902
D	48.841
E	232.341
F	82.527
other	63.266
total	1.888.907

Search Link

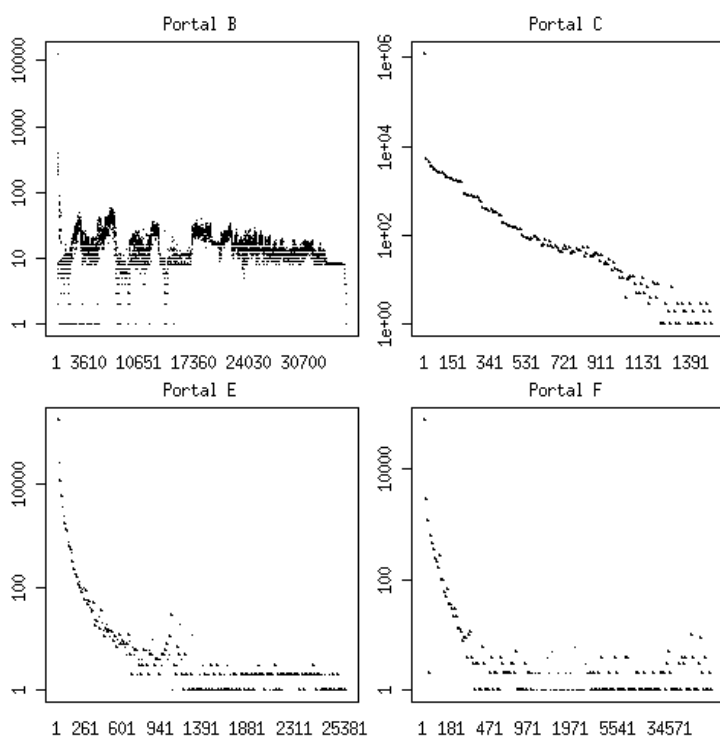
retrieve a first page resultset



- counting unique query occurrence
- start page, how much automatic?
- preset portal queries
`lom.educational.context=PO`

Pagination Links

retrieve another resultset page



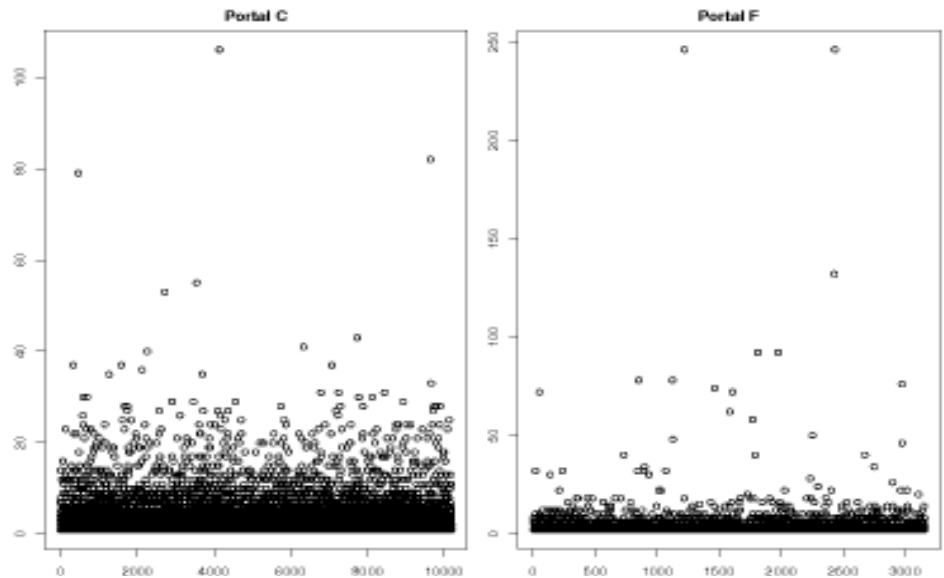
- mapping unique startRecord occurrence
- automatic querying behaviour
- normal querying behaviour

startRecord →

Result Links

retrieve a specific record

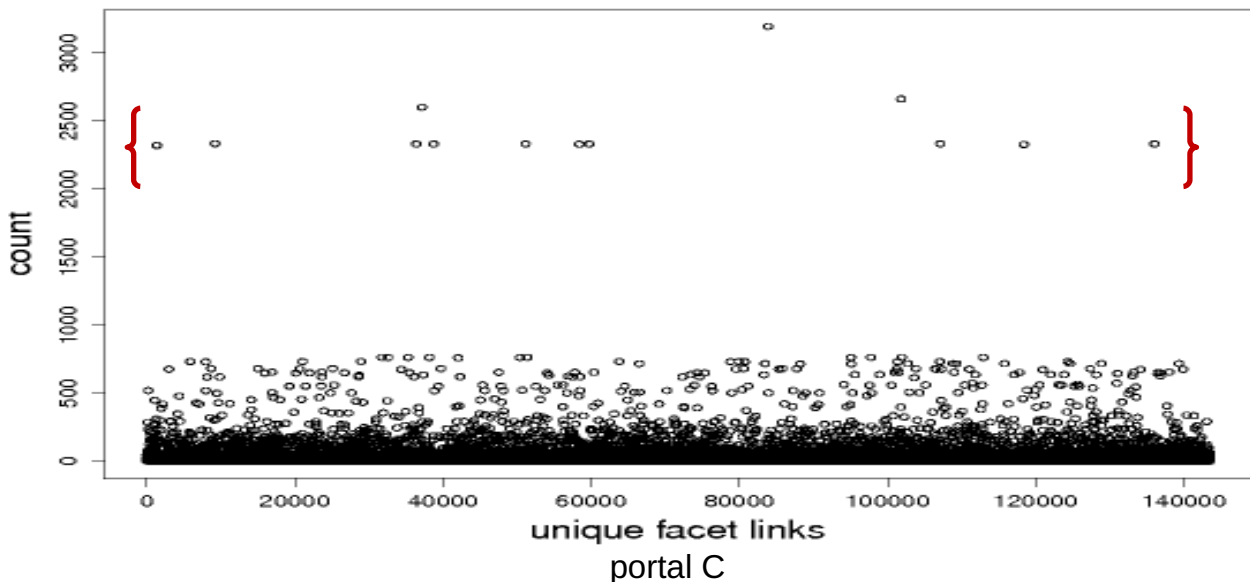
- not much result querying going on
- Edurep results are dynamic



Facet Links

retrieve the amount of records for that facet
(not using facet index)

- only portal c used this link type
- points corresponded with facet links on page



Observation Summary

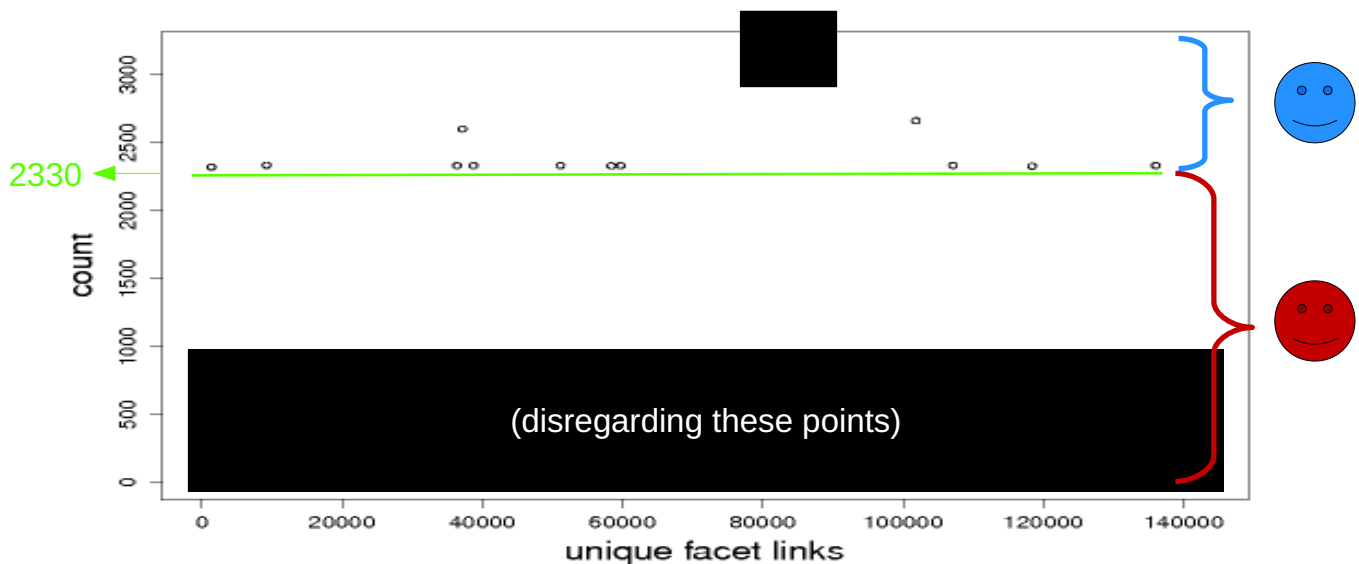
Automated querying indicated in:

- Facet links
- Pagination link

The question is how much...

Portal C's Facet Links

- human generated, but inefficient, distorting results
- 2330 x 12



Pagination Filtering

with maximumsRecord = 10
startRecord > 200 = resultpage 20

- Crude method:
startRecord > 200 = automated (PAG1)
- Elegant method:
filter pagination ranges (PAG2)

Pagination Ranges

first – 1 – 2 – 3 – 4 – 5 – last

1. query=dog&startRecord=1&maximumRecords=10
2. query=dog&startRecord=11&maximumRecords=10
3. query=dog&startRecord=21&maximumRecords=10
4. query=dog&startRecord=31&maximumRecords=10
5. query=dog&startRecord=41&maximumRecords=10

A Pagination Range

- is a set of equal queries, except for startRecord
 - minimum of 10 queries
 - maximum startRecord value > 200
 - a startRecord difference of maximumRecords between each query

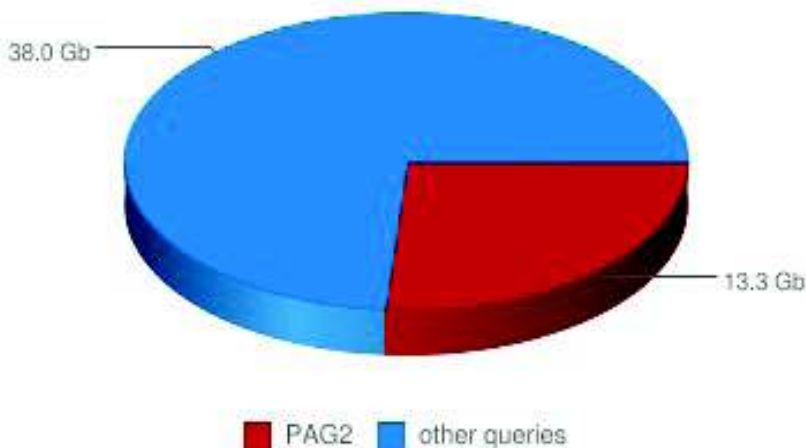
Pagination Results

portal	requests	crude (PAG1)	elegant (PAG2)
A	41.690	-15.237	-13.355
B	126.340	-105.026	-89.710
C	1.293.902	-15.255	-15.654
D	48.841	-47	-62
E	232.341	-1778	-1.815
F	82.527	-1.778	-205
total	1.825.641	-137.749	-120.801

- PAG2 did not take into account:
 - heads and tails outside the logs
 - first/previous/next/last page links

Results in Context

- PAG2's queries requested 13,3 Gb
- 6.6% of total requests



Immediate Findings

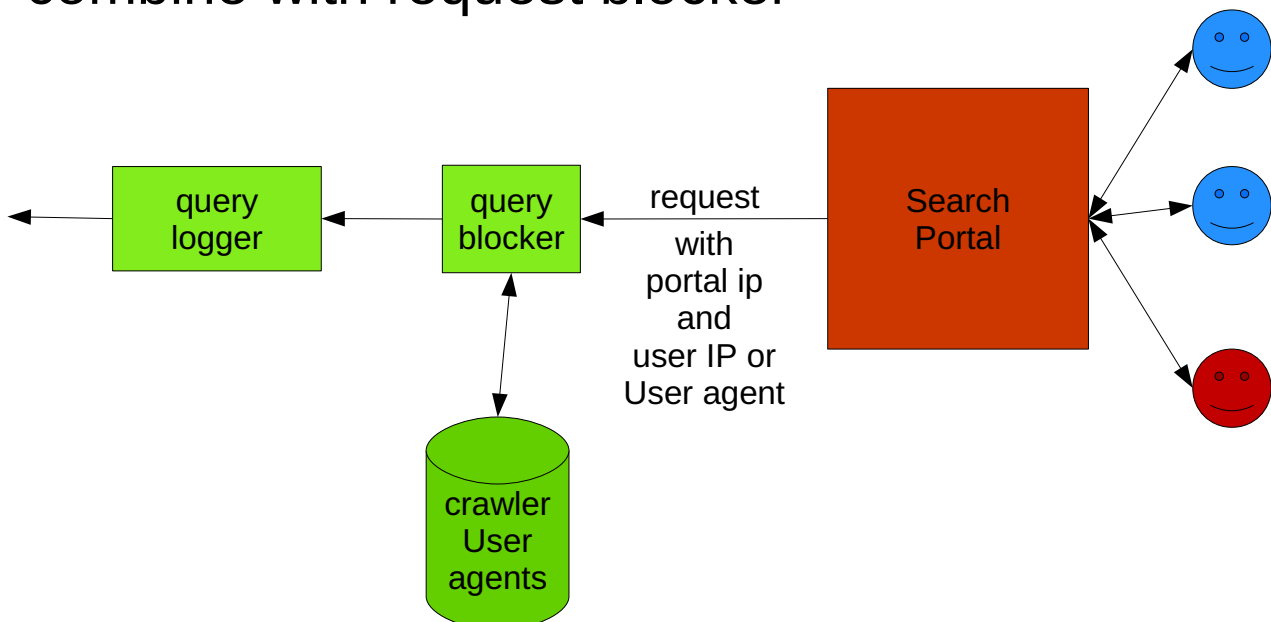
- Advise
 - blocking crawlers → Robot Exclusion Protocol
 - more efficient querying
- Administration
 - expanding & automating scripts
 - integrate in tooling

Research Improvement

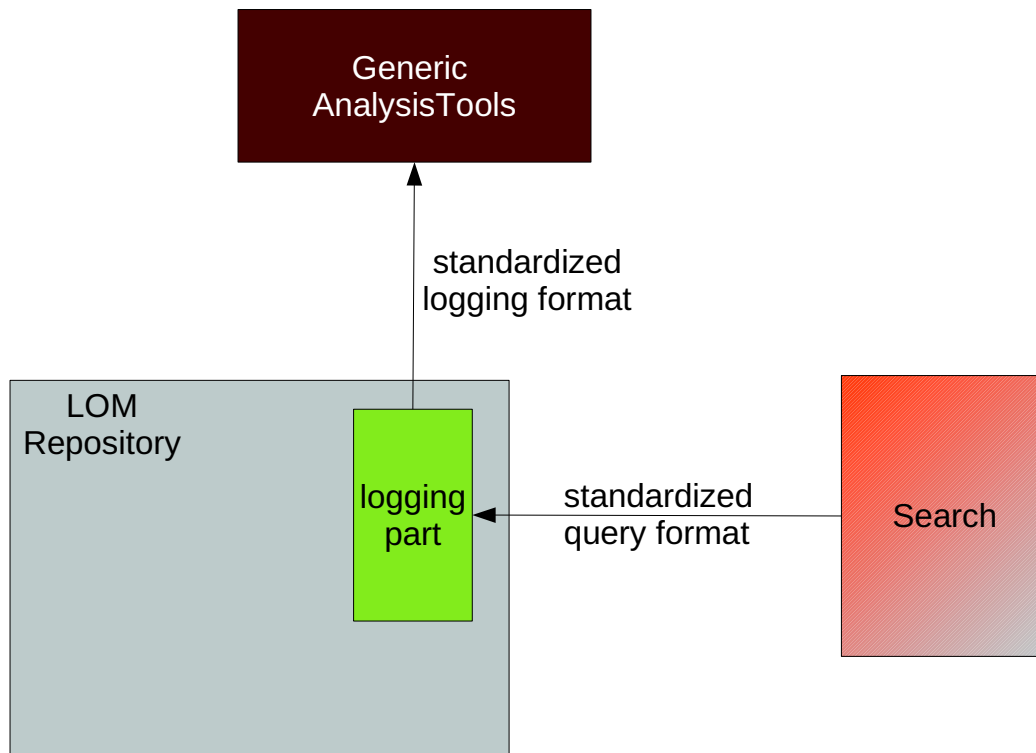
- use and compare with more logging
- detailed information on query
 - query={beer}AND{lom.general.title=abuse}AND{lom.technical.format=text/html}
- Use specific portal parameters in analysis
 - size and format of pagination
 - other search interaction links

Possible System Changes

- require user IP address or User agent in query
- combine with request blocker



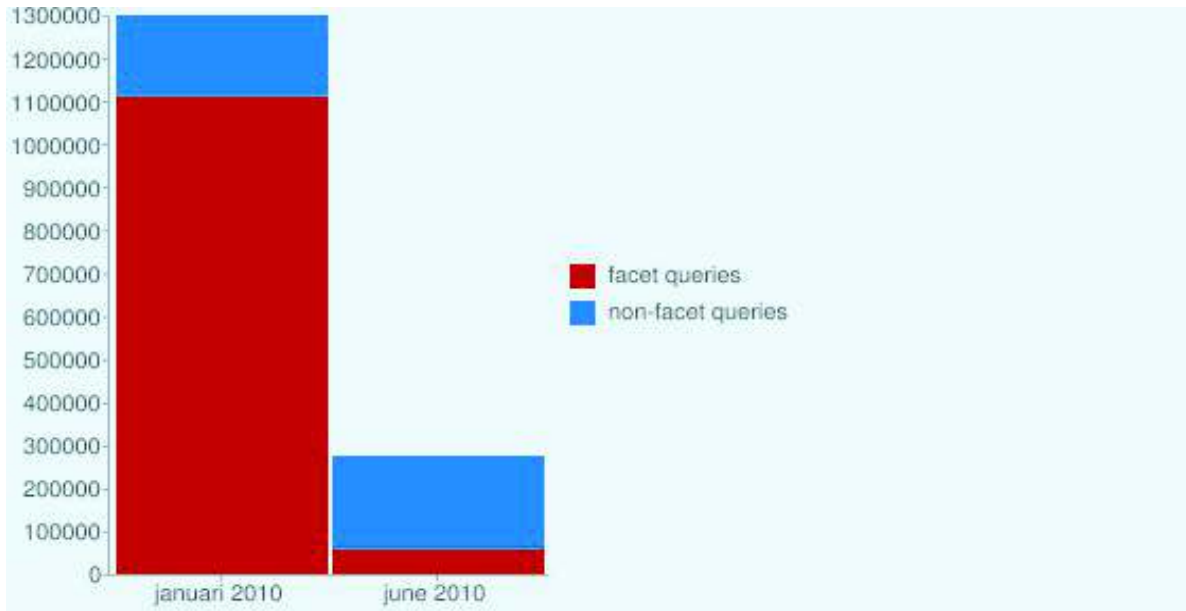
Generic Logging Tools



Some continued work...

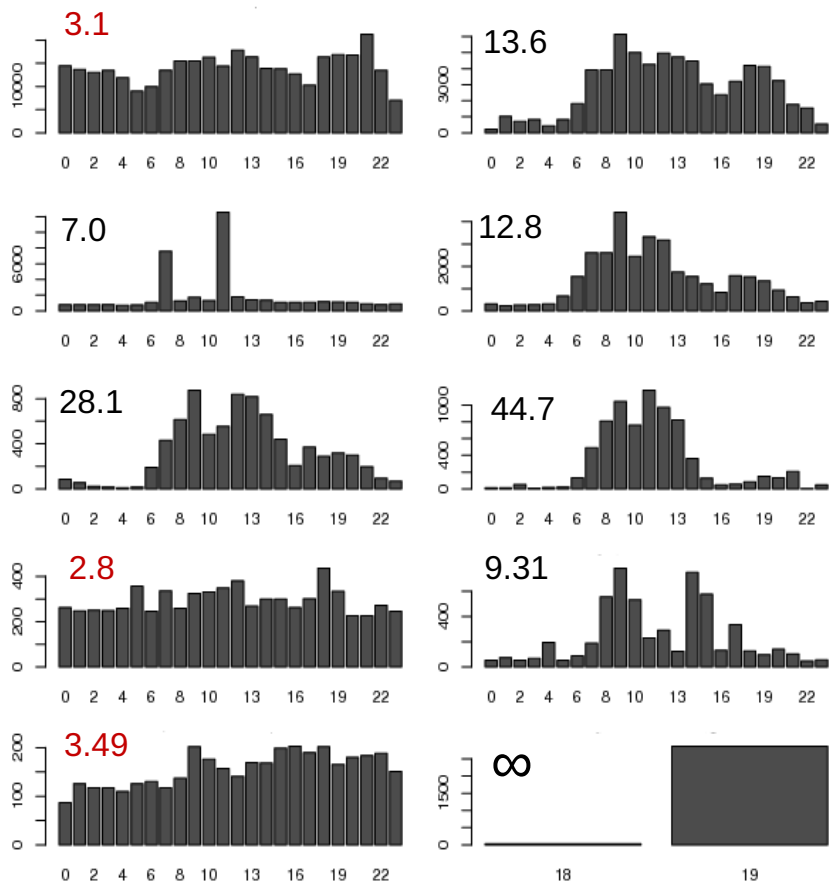
Advise Effect

- Results for consulting portal C on effective querying.

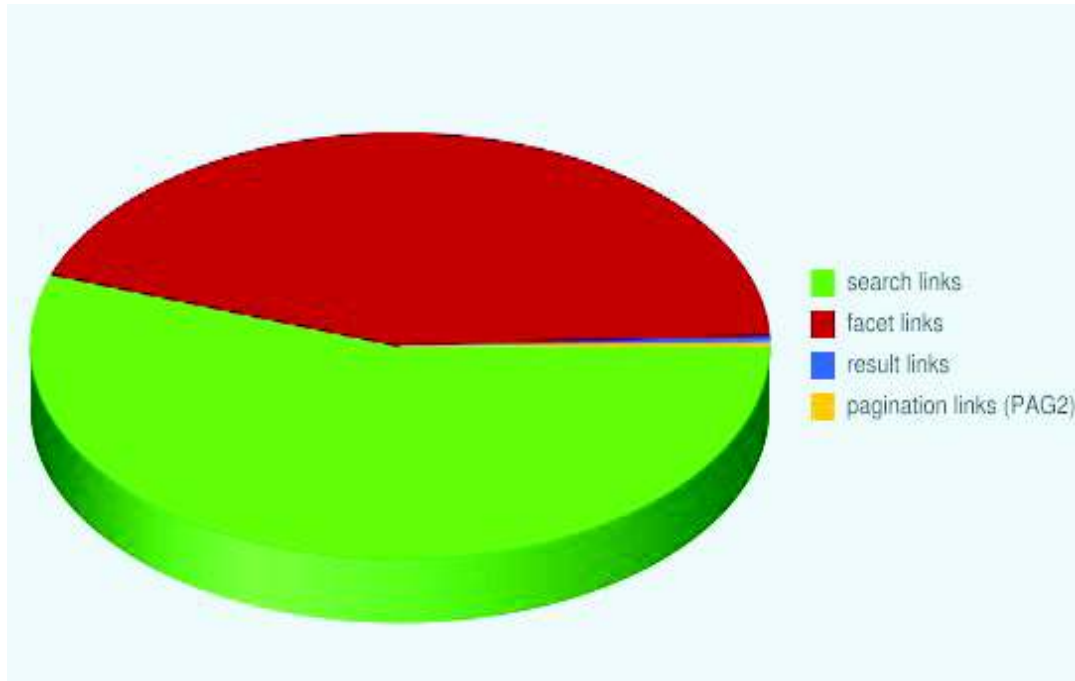


Querying, day and night

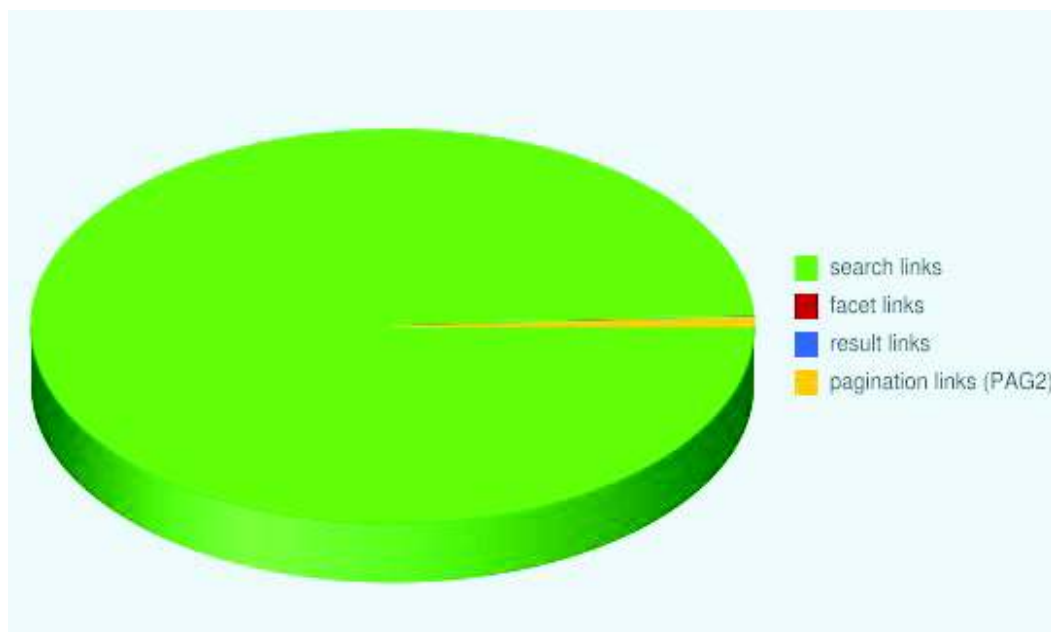
- Dutch content, so access from 1 timezone
- day when hour > 5:00 & < 0:00
- if ratio day / night < 5 automatic querying?



Combining Link Types Portal F



Combining Link Types Portal E



Combining Link Types Portal C

