Characterizing user groups of DBpedia-NL through user log analysis

Frank Walraven - 2542405
What is DBpedia?

- **Linked Data:**
  - “A method of publishing structured data so that it can be interlinked and become more useful through semantic queries” (Ismayilov & Kontokostas, 2017)

- **Freely available**
  - SPARQL-endpoints
  - Local dumps of DBpedia

- **Data from Wikipedia editions in 111 languages**

- **Users according to Official DBpedia Wiki:**
  - Large companies
  - Libraries
  - Researchers
  - Web developers.

- **Average of 7,343,939 hits per day (October 2016 - December 2017)**
Dbpedia-NL

- DBpedia is divided into local chapters per language
  - Coordinated by DBpedia Internationalization Committee
- Focus on DBpedia-NL
Problem

- Improving DBpedia-NL
  - Who are the users and what do they expect?
- Limited data available on users
Research questions

- Research questions:
  - What is a good way of characterizing user groups that use DBpedia-NL?
  - What is DBPedia-NL currently used for?
Method - Summary

- Data-driven method
- Validated through survey
- Data-driven method:
  - User log analysis in combination with local DBpedia-NL dump to retrieve the popular categories of DBpedia-NL
  - Subjects and hierarchical relations to characterize the log entries into grouped categories
- Validation through survey
  - Data-driven method results ↔ Survey results
- International version of DBpedia
Method - Pre-processing user log

● Pre:
  ○ xxx.xx.xx.xxx - - [07/Dec/2017:03:34:23 +0100] "GET /resource/Roger_Cicero HTTP/1.1" 303 -

● Post:

● Office Access 2016 issues
  ○ Apostrophes
  ○ Brackets
  ○ Minus
  ○ “GET”
  ○ HTTP
Method - IP-address analysis

- Classify most used IP-addresses
  - General description, not conclusive

<table>
<thead>
<tr>
<th>IP-address classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
</tbody>
</table>
Method - IP-address analysis

- IP-address lookup
  - Useful information:
    - Host
    - Country
    - IP owner info
    - Domain owner info
  - IP: 68.180.xxx.xx
    - Host: b110018.yse.yahoo.net
    - Country: United States
  - Not always possible to determine if bot or not
Results - IP-address analysis

- **Top 30 most used IP-addresses:**
  - 80% of IP-addresses Class A & bot
  - 16.67% unsure
  - 3.33% no bot

- **Excluding Class A from most used IP-addresses:**
  - 60% bots
  - 10% no bot
  - 30% unsure

<table>
<thead>
<tr>
<th>#</th>
<th>IP-address</th>
<th>Frequency</th>
<th>IP-class</th>
<th>Bot?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.180.xxx.xxx</td>
<td>1981562</td>
<td>A</td>
<td>Yes (Yahoo)</td>
</tr>
<tr>
<td>2</td>
<td>17.142.xxx.xxx</td>
<td>782012</td>
<td>A</td>
<td>Yes (Apple)</td>
</tr>
<tr>
<td>3</td>
<td>130.251.xxx.xxx</td>
<td>130780</td>
<td>B</td>
<td>Unsure</td>
</tr>
<tr>
<td>4</td>
<td>68.180.xxx.xxx</td>
<td>123090</td>
<td>A</td>
<td>Yes (Yahoo)</td>
</tr>
<tr>
<td>5</td>
<td>78.46.xxx.xxx</td>
<td>82204</td>
<td>A</td>
<td>Unsure</td>
</tr>
<tr>
<td>6</td>
<td>46.229.xxx.xxx</td>
<td>53190</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>46.229.xxx.xxx</td>
<td>51658</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>46.229.xxx.xxx</td>
<td>51314</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>46.229.xxx.xxx</td>
<td>51286</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>46.229.xxx.xxx</td>
<td>51192</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>46.229.xxx.xxx</td>
<td>51164</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>46.229.xxx.xxx</td>
<td>50690</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>46.229.xxx.xxx</td>
<td>50656</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>46.229.xxx.xxx</td>
<td>50414</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>46.229.xxx.xxx</td>
<td>50326</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>46.229.xxx.xxx</td>
<td>50146</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>17</td>
<td>46.229.xxx.xxx</td>
<td>44992</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>46.229.xxx.xxx</td>
<td>43850</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>138.201.xxx.xxx</td>
<td>41336</td>
<td>B</td>
<td>Unsure</td>
</tr>
<tr>
<td>20</td>
<td>46.229.xxx.xxx</td>
<td>41327</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>21</td>
<td>46.229.xxx.xxx</td>
<td>40930</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>22</td>
<td>202.180.xxx.xxx</td>
<td>32570</td>
<td>C</td>
<td>Unsure</td>
</tr>
<tr>
<td>23</td>
<td>46.229.xxx.xxx</td>
<td>26408</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>24</td>
<td>68.180.xxx.xxx</td>
<td>25803</td>
<td>A</td>
<td>Yes (Yahoo)</td>
</tr>
<tr>
<td>25</td>
<td>216.244.xxx.xxx</td>
<td>20600</td>
<td>C</td>
<td>Yes (Wowrack.com)</td>
</tr>
<tr>
<td>26</td>
<td>194.116.xxx.xxx</td>
<td>19904</td>
<td>C</td>
<td>Unsure</td>
</tr>
<tr>
<td>27</td>
<td>17.142.xxx.xxx</td>
<td>18258</td>
<td>A</td>
<td>Yes (Apple)</td>
</tr>
<tr>
<td>28</td>
<td>216.244.xxx.xxx</td>
<td>13996</td>
<td>C</td>
<td>Yes (Wowrack.com)</td>
</tr>
<tr>
<td>29</td>
<td>216.244.xxx.xxx</td>
<td>13054</td>
<td>C</td>
<td>Yes (Wowrack.com)</td>
</tr>
<tr>
<td>30</td>
<td>160.45.xxx.xxx</td>
<td>12520</td>
<td>B</td>
<td>No (Stanford researchers)</td>
</tr>
</tbody>
</table>
Method - Categories

- Wikipedia Categories are represented in DBpedia using DCMI terms and the SKOS vocabulary.
  - “Dcterm:subject”
    - “Android_TV” resource:
      - “category-en:Google”
      - “category-en:Android_(operating_system)_software”
  - Level 1 category
    - “Skos:broader”
      - Connect subcategories and supercategories
      - Level 2 category

Resource (URI Request)

Level 1 categories (dcterm:subject)

Level 2 categories (skos:broader)
Method - Unused data

- Not all resources are linked to Level 1 or Level 2 categories
- Unused data Level 1 categories:
  - 787,039 unique URI requests in user log
  - 33.79% (265,926 URI requests) unused
- Unused data Level 2 categories
  - 70,892 unique Level 1 categories
  - 6.72% (4,766 Level 1 categories) unused
Method - Level 1 categories

- Extract Level 1 categories
  - List of URI Requests and their corresponding Level 1 categories
    - SELECT level_1_categories.categories, uri_list.uri
      FROM level_1_categories INNER JOIN uri_list
      ON level_1_categories.uri = uri_list.uri;
    - Sorted using Office Excel 2016
  - Normalized frequencies
    - High in-degree of resources

\[
\text{Normalized Frequency} = \frac{\text{# of UNIQUE resources in user log per Level 1 Category}}{\text{Total # of linked resource in DBpedia–NL per Level 1 Category}}
\]
**Result - Level 1 categories**

- **IUCN-status_niet_bedreigd**
  - Rank 2
  - Normalized frequency: 27%

- **Dier_uit_het_Palearctisch_gebied**
  - Rank 10
  - Normalized frequency: 14%

<table>
<thead>
<tr>
<th>Top 10 Level 1 categories (unique resources per Level 1 category)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 category</strong></td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Amerikaanse_film</td>
</tr>
<tr>
<td>IUCN-status_niet_bedreigd</td>
</tr>
<tr>
<td>Dramafilm</td>
</tr>
<tr>
<td>Amerikaans_acteur</td>
</tr>
<tr>
<td>Nederlands_voetballer</td>
</tr>
<tr>
<td>Amerikaans_filmacteur</td>
</tr>
<tr>
<td>Amerikaans_televisieacteur</td>
</tr>
<tr>
<td>Komische_film</td>
</tr>
<tr>
<td>Pseudoniem</td>
</tr>
<tr>
<td>Dier_uit_het_Palearctisch_gebied</td>
</tr>
</tbody>
</table>
Method - Level 2 categories

- Extract Level 2 categories from Level 1 categories
  - SELECT Count(T.Field1) AS CountOfField1, T.Field2 FROM (SELECT DISTINCT user_log.ip, level_1_category_list.level1cat FROM level_1_category_list INNER JOIN user_log ON level_1_category_list.uri = user_log.uri GROUP BY level_1_category_list.level1cat, user_log.ip) AS T GROUP BY T.Field2;
  - DISTINCT for unique Level 1 categories per Level 2 category

- List of broader categories that are popular according to user logs
### Result - Level 2 categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film_naar_regisseur</td>
<td>941</td>
</tr>
<tr>
<td>Muziekalbum_naar_artiest</td>
<td>596</td>
</tr>
<tr>
<td>Burgemeester_van_een_voormalige_Nederlandse_gemeente</td>
<td>584</td>
</tr>
<tr>
<td>Nummer_naar_artiest</td>
<td>513</td>
</tr>
<tr>
<td>Plaats_in_India</td>
<td>399</td>
</tr>
<tr>
<td>Nederlands_burgemeester</td>
<td>369</td>
</tr>
<tr>
<td>Sport_naar_Nederlandse_gemeente</td>
<td>357</td>
</tr>
<tr>
<td>Gemeente_in_Noordrijn-Westfalen</td>
<td>315</td>
</tr>
<tr>
<td>Bouwwerk_naar_Nederlandse_gemeente</td>
<td>277</td>
</tr>
<tr>
<td>Rijksmonument_naar_gemeente</td>
<td>266</td>
</tr>
</tbody>
</table>
Method - Survey

- Target population:
  - Current users of DBpedia-NL
    - Difficult because finding out who users are is main problem
    - Published on DBpedia-NL website on May 18th, 2018
    - Spread over Twitter by DBpedia-NL chapter members

- Type of survey:
  - Self-administered questionnaire
  - Google Forms
  - No instructions given
  - 4 parts:
    - Respondent information
    - How do they use DBpedia-NL
    - What do they use DBpedia-NL for
    - Improvements
Method - Survey result analysis

- **Respondent information**
  - Sort of respondents
  - Whether they use DBpedia-NL or not

- **How do they use DBpedia-NL**
  - Usage frequency
  - Way of accessing DBpedia-NL
  - Validation of data-driven method

- **What do they use DBpedia-NL for**
  - Categories
  - Validation of data-driven method

- **Improvements**
  - Current issues
  - Satisfaction
Result - Survey: Respondent information

- 5 respondents
- Aged 22 years or older.
- Fields of work:
  - Information architecture
  - Information Technology Development
  - Data science and heritage
  - Semantic Technology Provider
  - Software Engineer (in cultural heritage)
Result - Survey: How do they use DBpedia-NL

- 40% uses downloaded datasets
  - Does not show up in user logs

![Bar chart showing access methods](chart.png)
Result - Survey: What do they use DBpedia-NL for

- Survey results:
  - Geography
  - Books / Writers
  - Sport
  - Movies
  - Music

- Data-driven results:
  - Dutch & International municipality information
  - Books
  - Sport
  - Domestic and international movies
  - Music
Discussion

- Unused data:
  - Unused data Level 1 categories:
    - 787,039 unique URI requests in user log
    - 33.79% (265,926 URI requests) unused
  - Unused data Level 2 categories
    - 70,892 unique Level 1 categories
    - 6.72% (4,766 Level 1 categories) unused

- Data-driven method skewed by bots

- Validation through survey:
  - More time
  - Visit conferences
Conclusion

- Research questions:
  - What is a good way of characterizing user groups that use DBpedia-NL?
    - Proposed method is useful, but hard to validate through survey
  - What is DBpedia-NL currently used for?
    - Categories:
      - Domestic & International movies
      - Music
      - Sports
      - Dutch & International municipality information
      - Books
  - High amount of bots
Questions?