Robot Detection for Usage Statistics

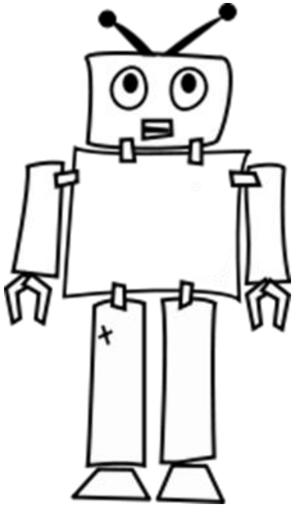
Short Presentation to introduce the breakout session “Robots”

Matthias Hitzler | hitzler@sub.uni-goettingen.de

Open-Access-Statistik

Overview

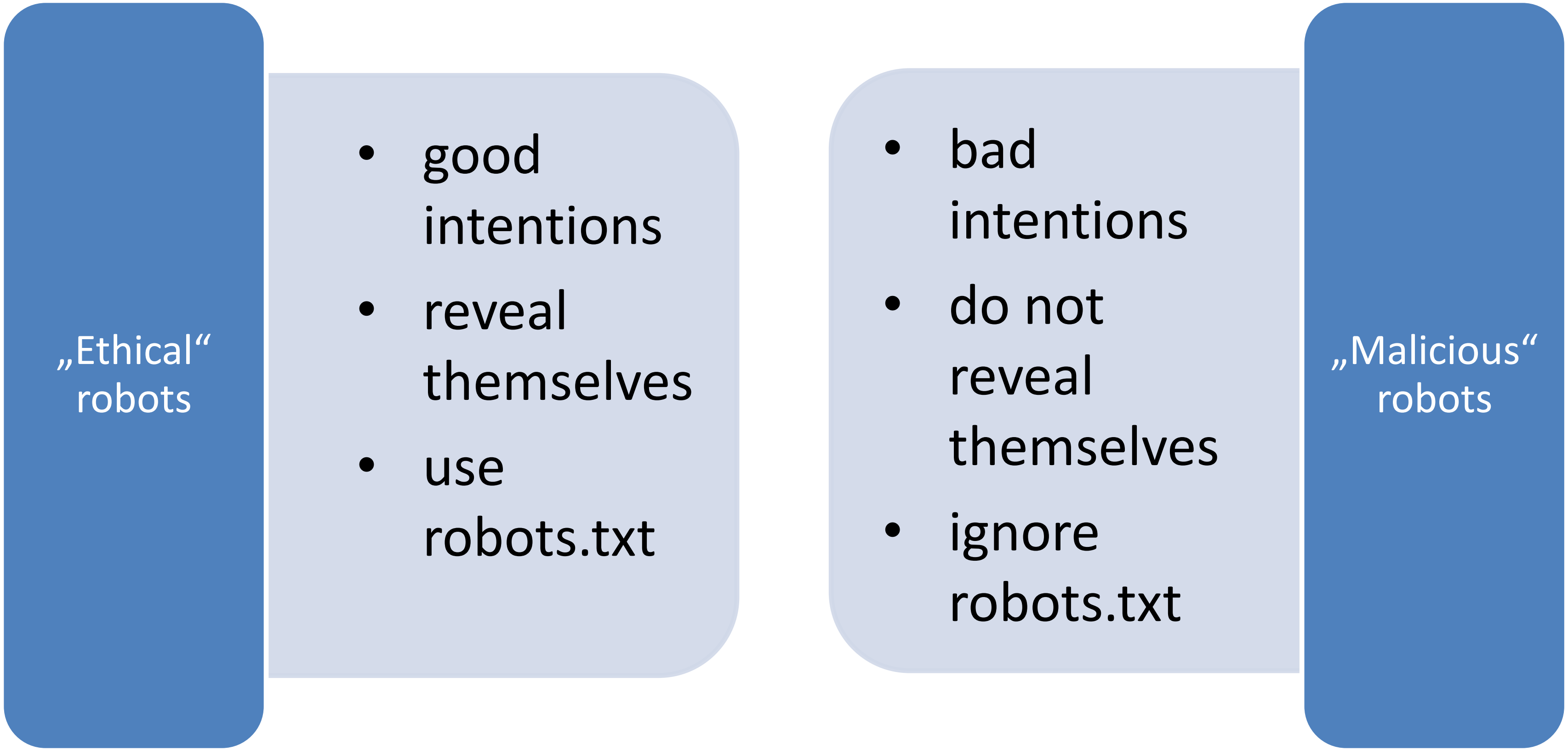
1. Robot definition
2. Identifying robots
3. Filtering robots in Open Access Statistics
4. A common list of robots
5. Suggestions for discussion



Robot Definition

“A robot is a program that automatically traverses the web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced.” (Martijn Koster 1999)

Robot Definition



„Ethical“ robots

- good intentions
- reveal themselves
- use robots.txt

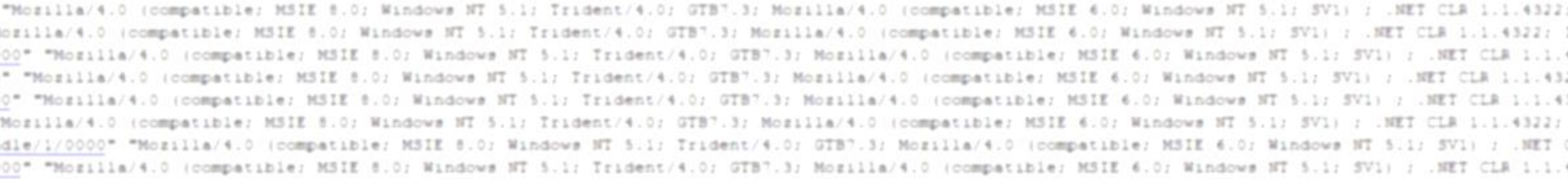
- bad intentions
- do not reveal themselves
- ignore robots.txt

„Malicious“ robots

Identifying Robots

Different approaches:

- IP adress
- User agent
- Heuristic methods
  - Time patterns
  - Document-related hits
  - Referer
  - Hit on robot.txt
  - Hit method



```
"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.4322;
ozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.4322;
00" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.
" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.43
0" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.4
Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.4322;
dle/1/0000" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET
00" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB7.3; Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) ; .NET CLR 1.1.
```

Filtering Robots in Open Access Statistics

In Open Access Statistics robots are identified by a list of user agents

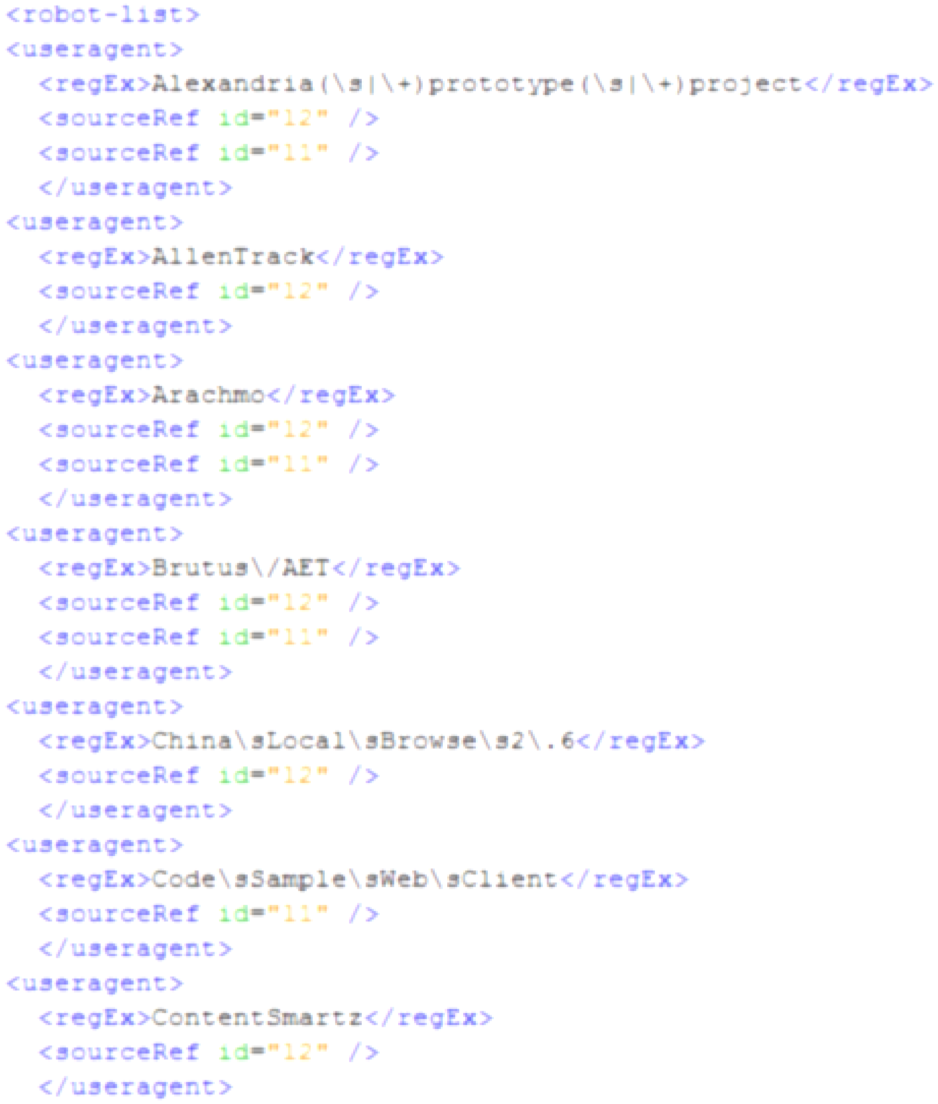
Current list contains entries from:

- COUNTER
- PLOS
- NEEO
- AWStats
- robottxt.org

A Common List of Robots

A way to go:

A common, freely available and regulary updated list of robot user agents



```
<robot-list>
<useragent>
  <regEx>Alexandria(\s|\+)prototype(\s|\+)project</regEx>
  <sourceRef id="12" />
  <sourceRef id="11" />
  </useragent>
<useragent>
  <regEx>AllenTrack</regEx>
  <sourceRef id="12" />
  </useragent>
<useragent>
  <regEx>Arachmo</regEx>
  <sourceRef id="12" />
  <sourceRef id="11" />
  </useragent>
<useragent>
  <regEx>Brutus\/AET</regEx>
  <sourceRef id="12" />
  <sourceRef id="11" />
  </useragent>
<useragent>
  <regEx>China\sLocal\sBrowse\s2\.6</regEx>
  <sourceRef id="12" />
  </useragent>
<useragent>
  <regEx>Code\sSample\sWeb\sClient</regEx>
  <sourceRef id="11" />
  </useragent>
<useragent>
  <regEx>ContentSmartz</regEx>
  <sourceRef id="12" />
  </useragent>
```

A Common List of Robots

KE Usage Statistics Guidelines:

1. It must be possible to 'timestamp' the list so that agents can refer to specific versions.
2. The list must be in a machine-readable format, and preferably in XML.
3. The extended list must be approved by COUNTER. Institutions that make use of the extended list should also be able to pass the COUNTER audit.
4. It must be possible to indicate the 'source' of each entry in the list (e.g. "COUNTER", "AwStats", "Plos", etc.)
5. It must be possible to access the robot list on the basis of a persistent URI.
6. It must be possible to manage different versions of the robot list. The most recent version must always be available through a uniform URL.

A Common List of Robots

Following points could be added to this list:

1. The list can and should be extended.
2. New versions of the list will be published on a regular basis.
3. As well as a timestamp, the list should also contain a version number to make it easier to identify and distinguish between different versions.
4. The basic list consists of the COUNTER, PLOS and NEEO lists.
5. New entries can be submitted by general public.
6. New entries are peer reviewed before being added to the list.
7. The continued availability of the list must be guaranteed.

Suggestions for Discussion

- Is such a regularly updated robot list desirable?
- How often should the list be updated?
- Who is permitted to add entries to the list?
- Which criteria for detecting and adding to the list must be applied?
- What importance have heuristic methods for detecting robots? How should they be added for usage statistics?