eGOVMON

eGovernment Monitor

A solution to the exact match on rare item searches

Morten Goodwin
Tingtun AS, Aalborg University

International Conference on Web Intelligence, Mining and Semantics (WIMS)
Sogndalen, Norway, 2011-05-27

Project no.: Verdikt 183392/S10

A solution to the exact match on rare item searches

- A approach/tool for locating eGovernment services
- Where does it fit in. What can it be used for?
  - Egovernment Surveys, Automatic measurements of web sites
- Problem definition
  - What is the exact match on rare item searches.
- Algorithms
- Results
  - Syntethic environments
  - Real web sites
- Conclusion and further work.

eGOVMON

eGovernment

- EGovernment means the use of IT to improve governments. E.g.
  - Make information available online to reach more citizens.
    - Local government budget, municipal calendar, and so on.
  - Make governmental services available online using web technologies.
    - Tracking of building permissions, mail records, interactive meetings, and so on.

Egovernment Measurements

- Benchmark eGovernments is common.
  - In Norway: DIFI/Norge.no, Consumers Council
  - International: UNPAN, Capgemini, Brown University, ...
- Most common to check the supply side of governments:
  - Web site, e-mail, mobile functionality, and so on.
  - Checking the supply side is easiest because it is available.
  - The supply side is what citizens use and is therefore the most important for the citizens.

Characteristics of eGovernment testers

- Fall into at least one of the three categories:
  - Big organisations: UN, Capgemini, and so on.
  - Focus on a specific eGovernment topic.
  - Focus on a small geographical or political area.
- The others have too few resources.

Manual assessment (1)

- Much of the work is assessed manually.
  - Existence tests:
    - An expert checks whether a service is available online.
    - Since interpretations may differ from expert to expert, and day to day, the same tests need to carried out many times.
  - Findability tests:
    - Can the service be found by actual users.
    - Representative users try to locate information within a time frame.

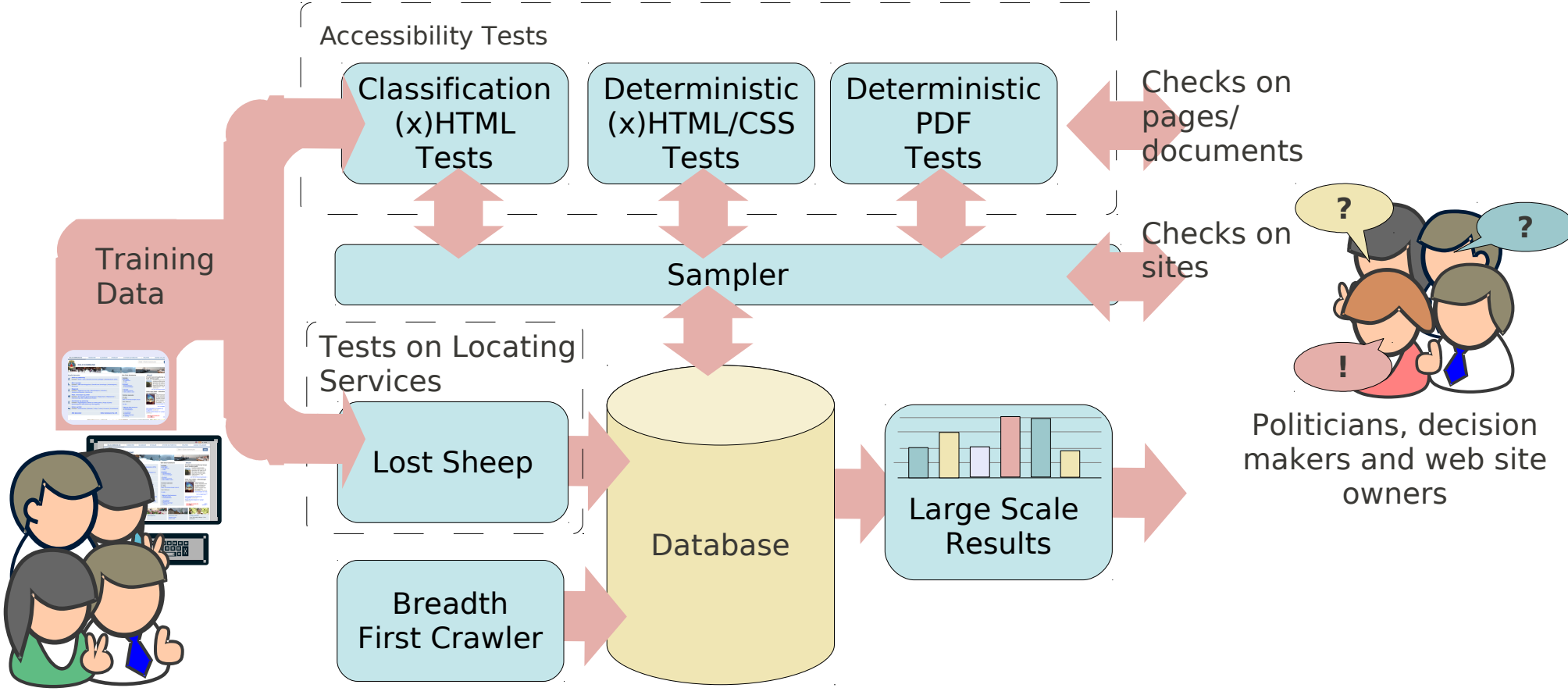
Manual assessment (2)

- Disadvantages:
  - Time consuming, costly, biased, in-frequent, not on demand,
- Advantages:
  - Is better at judging than machines, high accuracy, more manual tests available,
- It therefore makes sence to automize as much of the assessment as possible.

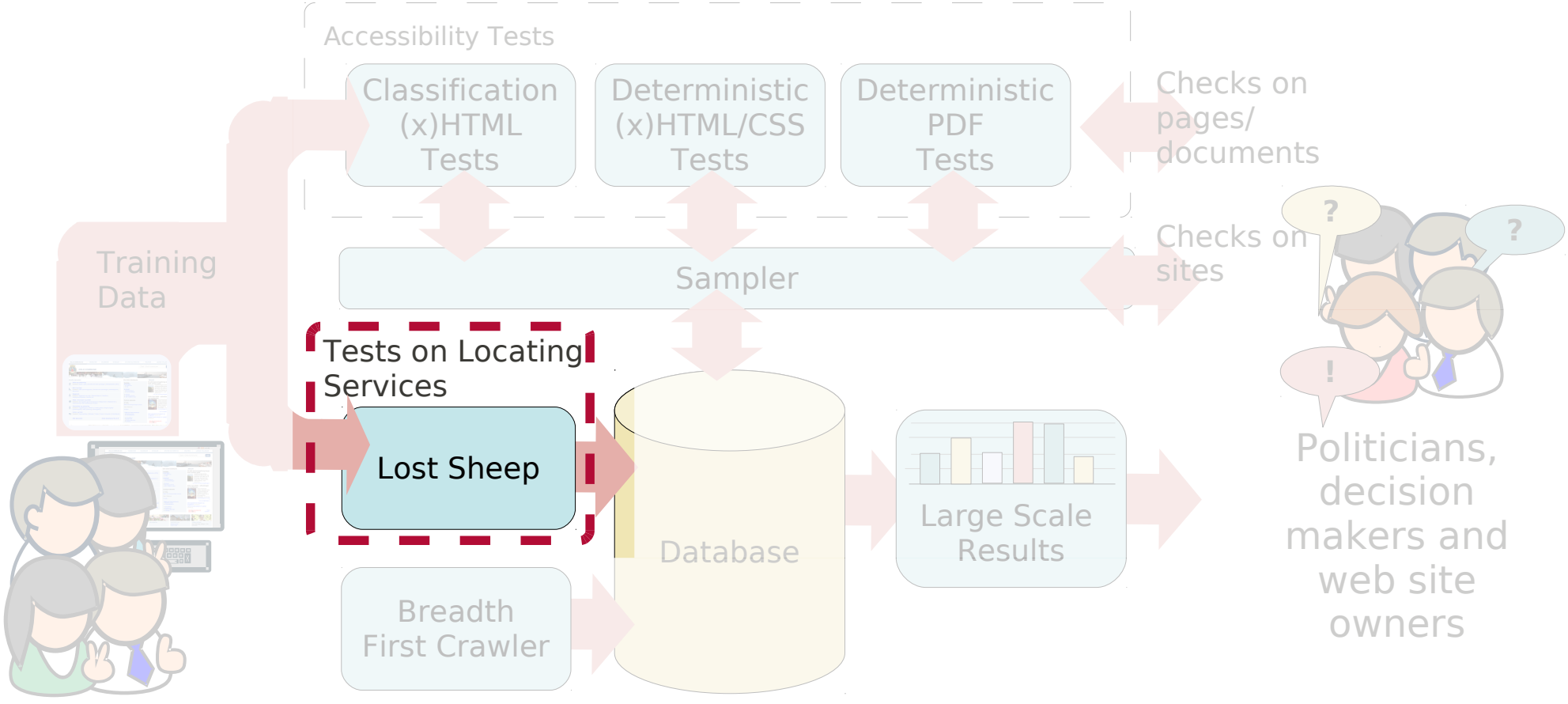
Motivations

- Motivations for eGovernment:
  - Introducing IT to improve governments.
- Motivations for this approach:
  - Introducing automation to improve eGovernment measurements.

Automatic testing (1)



Automatic testing (2)



Characteristics of services / information online (1)

- In most cases it is only one per web site.
  - E.g. The most recent budged.
- Restricted by robots.txt
  - Much information not available at main search engines.

Characteristics of services / information online (2)

- Government web sites are complex:
  - Many services available from different vendors, often available at the vendors web site.
    - Domain name not sufficient to describe web site.
- Government web sites link almost exclucive to their own information/ services:
  - No municipality link to the contact information of others.

Problem definition

- Find the one web page within a web page within the web site that matches the criteria, if any.
- Formally:
  - Let p_t be an unobservable target page.
  - Select a page in web sites S so that it is expected to be the p_t while minimizing the number of downloaded pages.
- Most useful when web sites are large.
- **E**xact **M**atch on **R**are **I**tem **S**earch (EMRIS).
- Unlike similar problems: You can easily decide the correct target page prior to running the algorithm
  - No (subconsciously) favoring of their own algorithm.

Related algorithms (1)

- EMRIS
  - Similarity search: Follow links to pages which are most similar to the training data using cosine similarity.
    - Requires download of pages before choosing if it should be followed.
  - Degree based search: Follow links to pages which has many links.
  - SIMDEG: Merging of Similarity and Degree based search

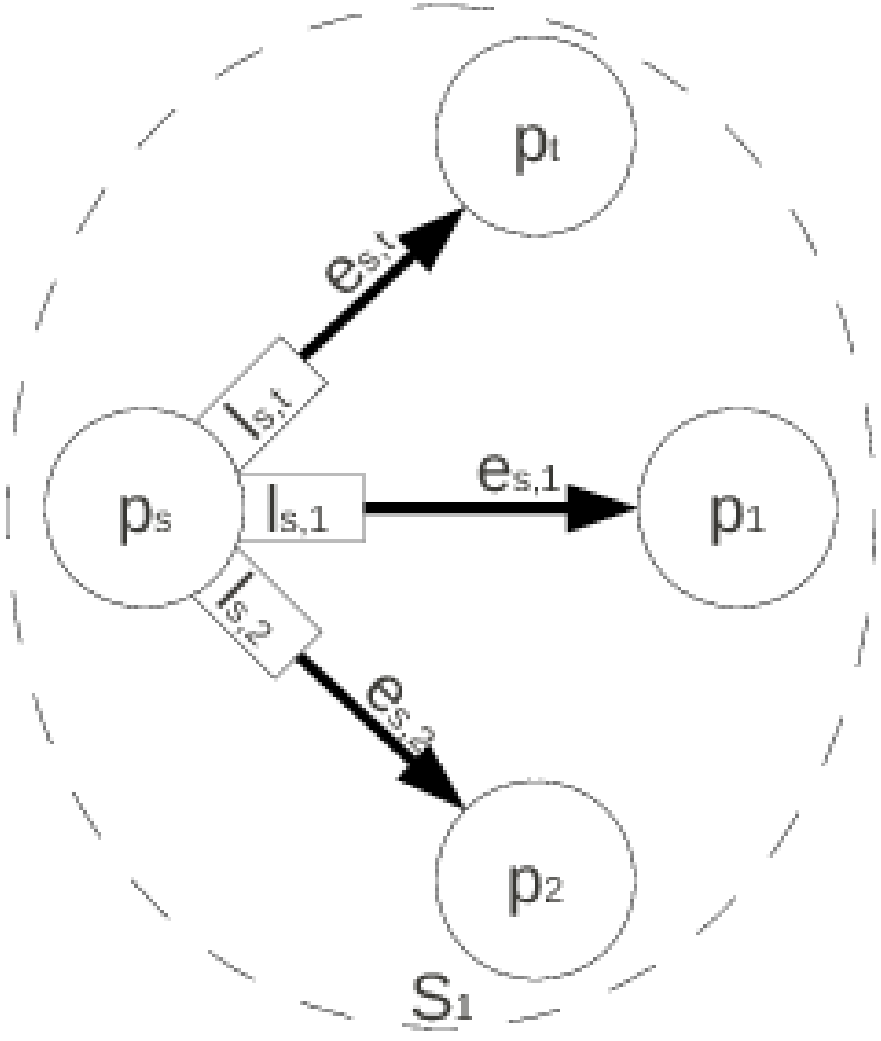
Related algorithms (2)

- Focused crawlers (e.g. Fish search, shark search):
  - Not focused on finding one page.
- Web page classifiers:
  - Not minimizing the number of pages to be evaluated.
- Search algorithm (e.g. A*):
  - The target page is observable, which is a a different problem.
- Lost sheep: Not a replacement

Modelling the web site

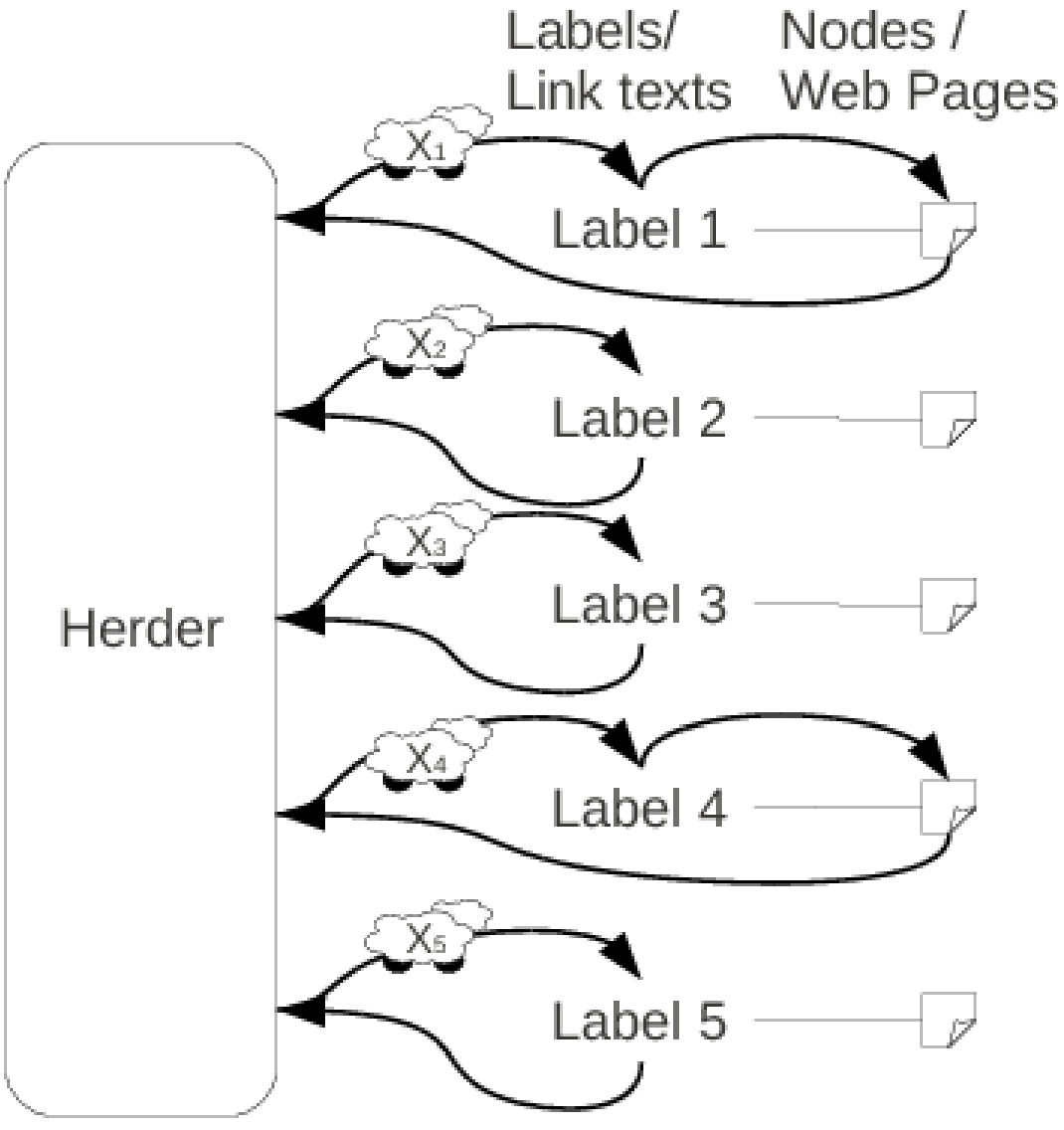
- Web site: A labled directed graph G(V,E,L)
  - V = pages
  - E = links
  - L = labels / link texts

Web site example (1)



- p_s: Starting page.
- p_t: Target page with contact information
  $l_{s,t}$: Contact us
- p_1: General information
  $l_{s,1}$: About us
- p_2: Image gallery
  $l_{s,2}$: Image gallery

Lost Sheep


Labels/
Link texts
Nodes /
Web Pages
Herder
X1
Label 1
X2
Label 2
X3
Label 3
X4
Label 4
X5
Label 5

How is lost sheep an improvement?

- Uses link texts as a pre-classifier.
  - Downloads fewer pages and makes the classification problem easier.
- Can use any classifier.
  - Use the classifier that fits the problem to be solved.
  - Not intended to replace existing classifiers.
- Works with web sites which are not formally scoped.
- Is able to locate the target page than comparable algorithms with a higher accuracy and fewer downloaded pages.

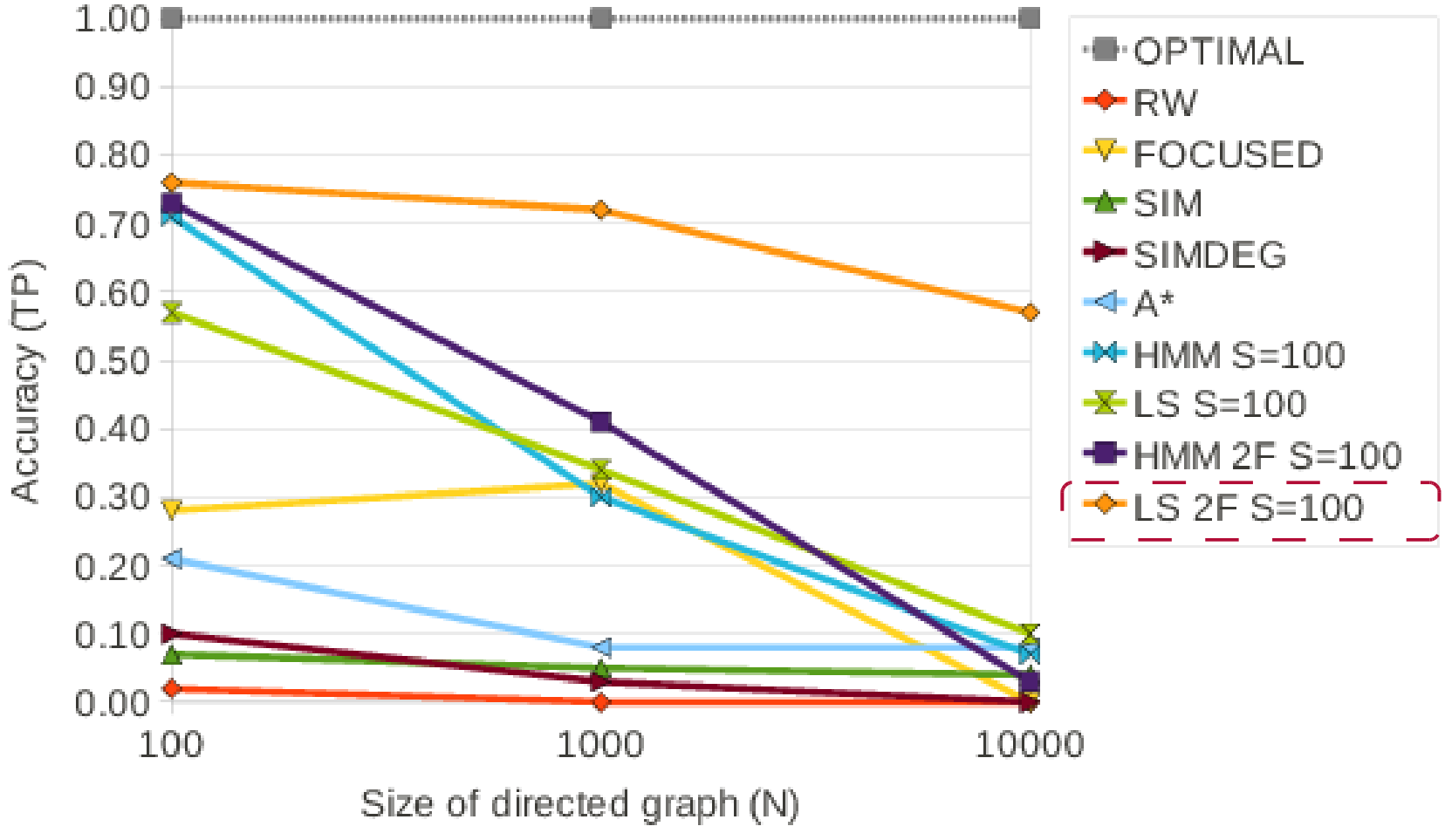
Practical implementation notes

- Stemming is applied.
- Stop words removed.
- Each sheep is:
  - (1) A learning automata classifier, including using tf-idf.
  - (2) Cosine similarity approach.
- The sheep stop when:
  - Confidence: > 0.75 (tried many)
  - Maxdepth: 5

Synthethic environment

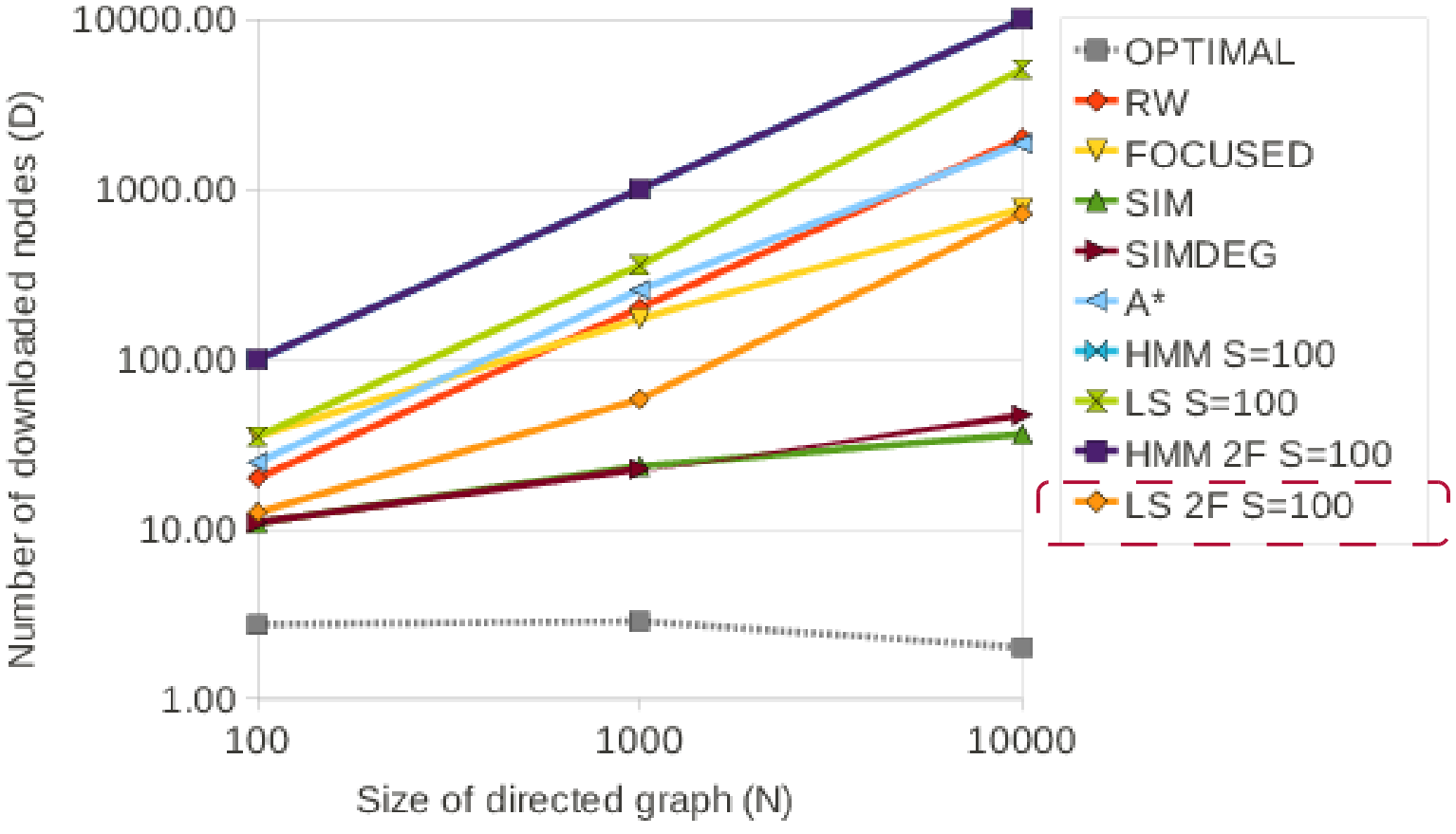
- Target page, p_t, is chosen at complete random
- Training data: 75 % of words randomly chosen words within p_t 25% outside

Synthethic Environment Results - Accuracy



eGOVMON

Synthethic Environment Results – Number of downloads



Real environment tasks (1)

- 13 realistic tasks in 427 Norwegian local government web sites.
  - Finding at most one page with information or services on a web site.
    - E.g. The latest anual budget.
- Locating transparency information and services in local government web sites.
  - Based on commonly assessed information and state of the art.
  - Open government data.
- Two tasks completely assessed manually. 11 tasks assessed 10%.
  - The target page chosen **before** the algorithm.

Real environment tasks (2)

1) Contact information
2) Recent information section
3) Budget
4) Local government calendar
5) Local government plan
6) Zoning information and plans
7) Mail record
8) Search functionality (on a single page)
9) Online local government board meetings
10) Online local government executive council
11) Chat with administrative or political officials
12) Video of city or municipality board meetings
13) Online city or municipal plan meeting

Main references: Sundance, eGep, Leapfrog, Norge.no, Local Government Stakeholders and decision makers.

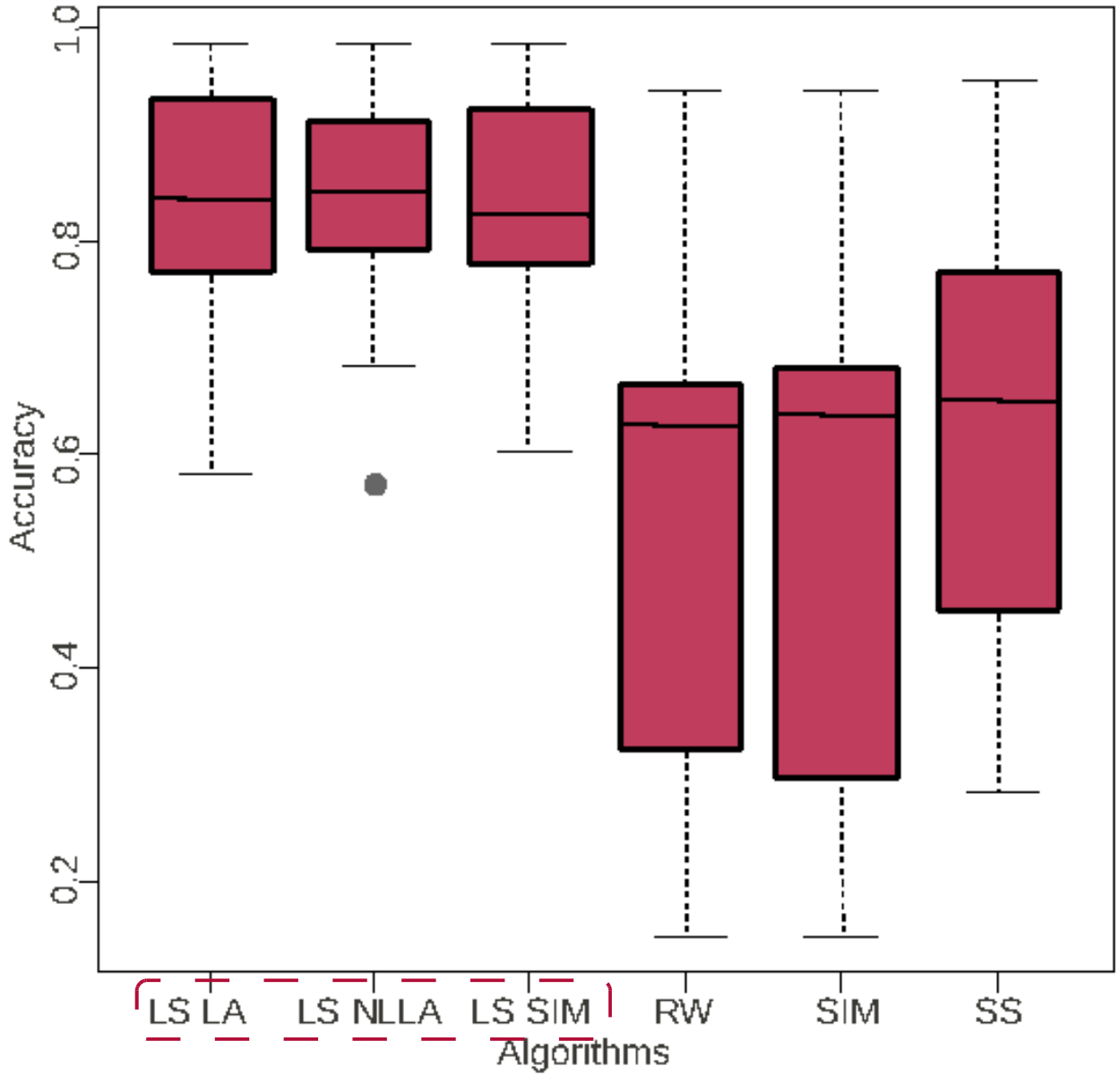
Results

Task	Existance	Findable
Contact information	416	345
Recent information section	350	168
Budget	199	60
Local government calendar	238	105
Local government plan	216	65
Zoning information and plans	155	33
Mail record	379	311
Search functionality (on a single page)	364	179
Online local government board meetings	332	143
Online local government executive council	292	102
Chat with administrative or political officials	20	9
Video of city or municipality board meetings	27	14
Online city or municipal plan meeting	39	13

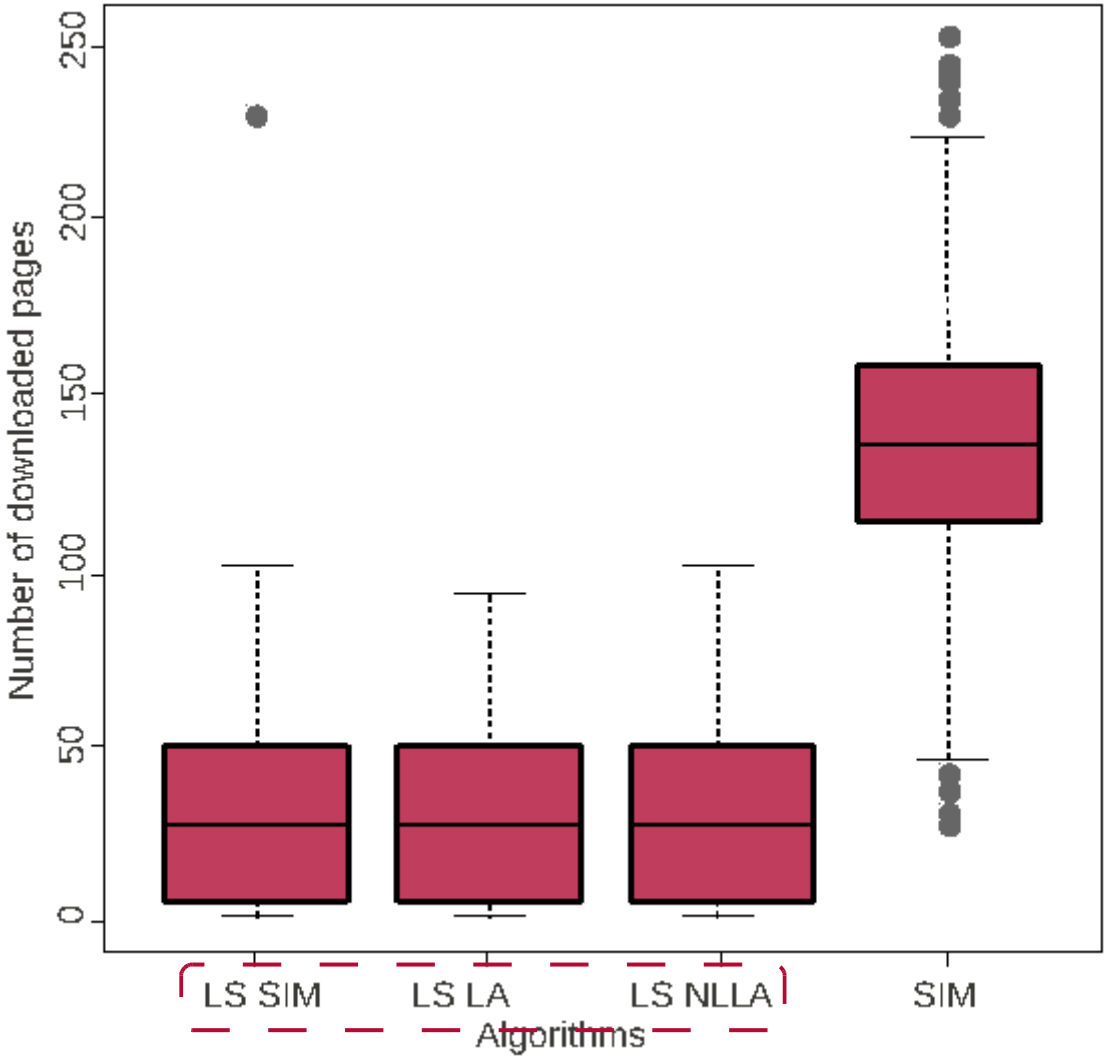
Sample size: 427 web sites.
Training data: Between 10-20 pages.

eGOVMON

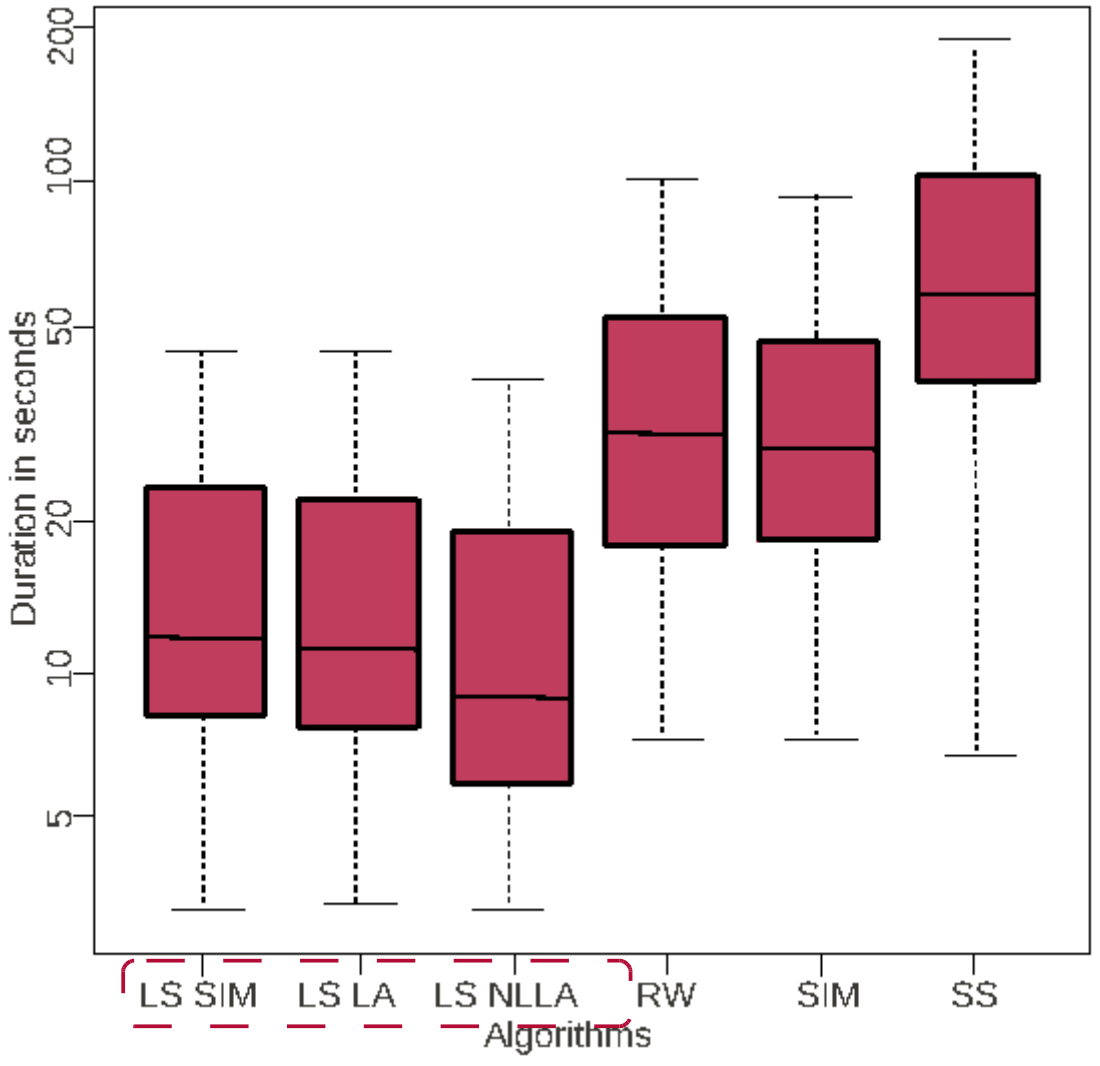
Real Environment - Accuracy



Real Environment – Number of downloaded pages



Real Environment - Duration



eGOVMON

Usefullness of the results

- Reports
  - Existsance (if a target page exists) and
  - Findability (an indication if a target page can be found by real users) based on number of clicks.
- Depending on the resources available, the lost sheep could either be used:
  - Directly.
  - As input to experts for manual verification.
  - Prioritising of which tests should be carried out as user testing.

Further work

- Language independence.
  - Large eGovernment surveys are for many countries and languages.
  - Either language independance or classify multiple languages.
- Hands on experience on eGovernment surveys.

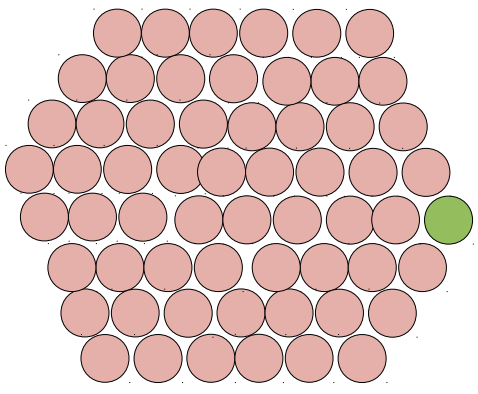
Thanks

Supplementary slides

Supplementary slides

Metrics (1) (Why TP is enough)

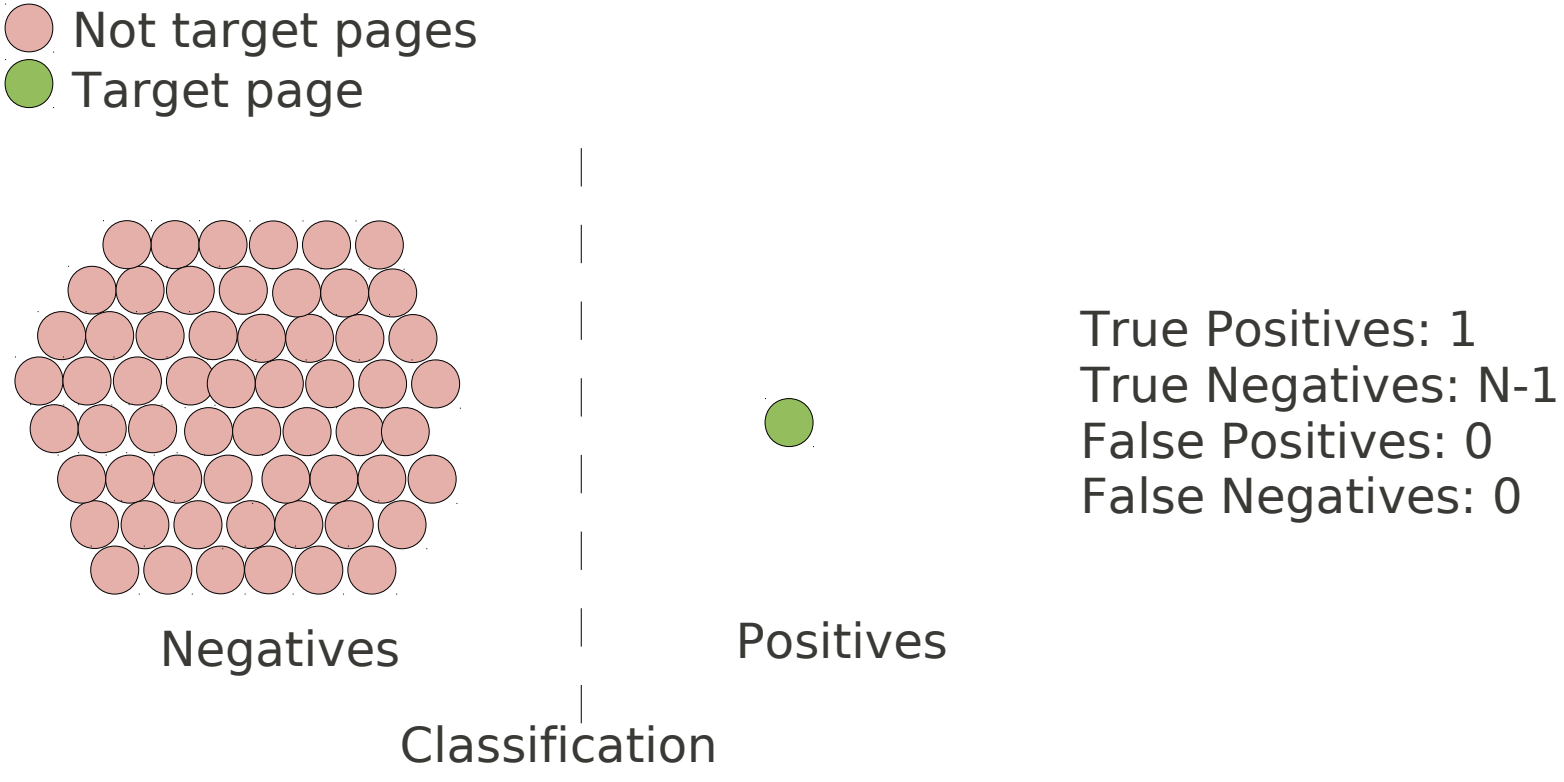
- Web site
  - Not target pages
  - Target page



Web Site

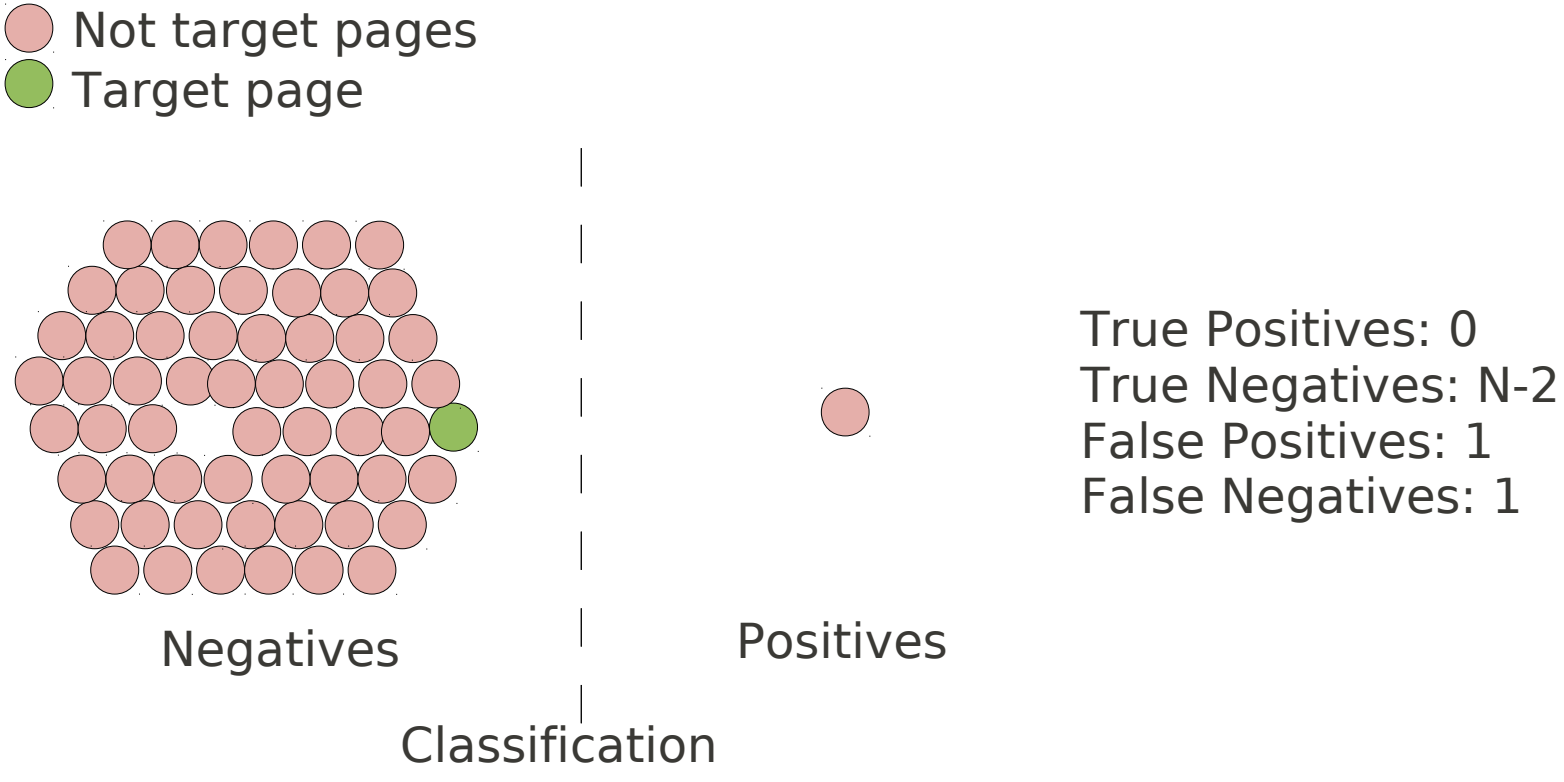
Metrics (2)

- Option 1: Has found the correct page.



Metrics (3)

- Option 2: Has found the correct page.



Metrics (4)

- Conclusion:
  - True Positive, True Negative, False Positive, False Positives can all be calculated if the following is known
    - N.
    - A TP is known.
  - Hence: If(TP==1):
    - TP=1, TN=N-1, FP=0, FN=0
  - Else:
    - TP=0, TN=N-2, FP=1, FN=1

Lost Sheep Algorithm (1)

Algorithm 4.1 Lost Sheep

1: $depth \leftarrow 0$
2: $q \leftarrow$ training data
3: $x'.confidence \leftarrow 0$
4: $p_i \leftarrow p_s$
5: $Visited \leftarrow$ empty set.
6: **while** $x'.confidence < threshold$ and $depth < maxdepth$ **do**
7: Add p_i to $Visited$.
8: Let $\mathbf{E} = \{e_{i,0}, \ldots, e_{i,n}\}$ be the n edges from p_i.
9: Let $\mathbf{L} = \{l_{i,0}, \ldots, l_{i,n}\}$ be the labels connected to each edge in $\mathbf{E}$.
10: The herder releases n sheep $X = \{x_0, \ldots, x_n\}$ where each sheep $x_j \in X$ is connected to edge $e_{i,j}$ if $p_i \notin Visited$.
11: **for all** $x_j \in X$ **do**
12: Let $l_{i,j}$ be the label connected to edge $e_{i,j}$.
13: Let p_j be the not yet visited page available from $e_{i,j}$.
14: $x_j.confidence \leftarrow 0$
15: $x_j.page \leftarrow p_j$

Lost Sheep Algorithm (2)

16: $tp'(x, o, q) \leftarrow$ a classifier, e.g algorithm 4.2, 4.3 or 4.4
17: $x_j \leftarrow tp'(sheep\ x = x_j, object\ o = l_{i,j}, training\ data\ q = q)$
18: **if** $x_j.shouldcontinue$ **then**
19: Download p_j
20: $x_j \leftarrow tp'(sheep\ x = x_j, object\ o = p_j, training\ data\ q = q)$
21: Add p_j to $Visited$.
22: **end if**
23: **end for**
24: $depth \leftarrow depth + 1$
25: Find $x'' \in X$ where $x''.confidence \geq x_j.confidence$ for all $x_j \in X$
26: **if** $x''.confidence > x'.normconfidence$ **then**
27: $x' \leftarrow x''$
28: $p_i \leftarrow x'.page$
29: **end if**
30: **end while**