Automatic retrieval of web pages with standards of ethics and trustworthiness within a medical portal.
What a page name tells us

Arnaud Gaudinat, Natalia Grabar, Celia Boyer
first.second@healthonnet.org, Health on the Net Foundation, SIM/HUG, Geneva, Switzerland

Problems
Large quantity of health information online
Uneven quality of health information online

Objectives
1. Help citizens and professionals
   finding reliable health web sites
2. Help HONcode reviewers
   analysing and accrediting health web sites

Context
HONcode of ethical principles

1. Authoritative: Indicate the qualifications of the authors
2. Complementarity: Information should support, not replace, the doctor-patient relationship
3. Privacy: Respect the privacy and confidentiality of personal data submitted by users
4. Attribution: Cite the source(s) of published information, data and medical and health pages
5. Justifiability: Site must back up claims relating to benefits and performance
6. Transparency: Accessible presentation, identities of editor and webmaster, easy email contact
7. Financial disclosure: Identify funding sources
8. Sponsorship: Clearly distinguish advertising from editorial content

Accreditation process at the HON Foundation
Web site submits a request to get the HONcode accreditation
Web site fills an online questionnaire
HON reviewers deeply analyse the web site
According to findings, the web site is
- Accredited
  - Accredited under condition to modify/add some information
  - Not accredited
Each web site is reviewed each year
HONcode consists of 8 principles
HONcode base covers over 5,500 web sites

Material
URL address
Content excerpt

Size of corpora

<table>
<thead>
<tr>
<th>Principle</th>
<th>Meaning</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Authoritative</td>
<td>2843</td>
</tr>
<tr>
<td>H2</td>
<td>Complementarity</td>
<td>2470</td>
</tr>
<tr>
<td>H3</td>
<td>Privacy</td>
<td>2974</td>
</tr>
<tr>
<td>H4</td>
<td>Reference</td>
<td>1855</td>
</tr>
<tr>
<td>H5</td>
<td>Justifiability</td>
<td>460</td>
</tr>
<tr>
<td>H6</td>
<td>Transparency</td>
<td>2593</td>
</tr>
<tr>
<td>H7</td>
<td>Financial disclosure</td>
<td>2088</td>
</tr>
<tr>
<td>H8</td>
<td>Sponsorship</td>
<td>1645</td>
</tr>
<tr>
<td>H9</td>
<td>Date</td>
<td>1645</td>
</tr>
</tbody>
</table>

Pre-processing of URLs
- Segmentation on non alphanumeric characters (\-?+/\)
- Distinction of significant parts of URLs
  - www.mettenchildcare/general/privacy.html
  - [user]/[content]-----

Sets of data
- Inurl/Outurl parts of URLs
- complete/reduced sets of principles
  reduced (-H5, -H9)

Learning platform
(Williams & Calvo, 2002)
Algorithms: Naive Bayes, SVM, kNN, DT
NB (Naïve Bayes)
SVM (Support Vector Machine)
kNN (k-Nearest Neighbour)
DT (Decision Tree)
Features: n-grams of words, n-grams of characters
Learning 90%, test 10%

Discussion & Perspectives
Three sets of results with URL-based categorisation
- inURL: less discriminant part of URLs
- SVM: better precision
- NaïveBayes: better recall
Content-based categorisation:
similar performances of algorithms
better results comparing to URL-based categorisation
Combination of URL and content-based categorisation
Multifold cross-validation

Quality categorisation
(HONcode principles detection)

Integration within search engine

Learning process
Language model
Crawling
Categorisation of pages
HONcode DB
User

Web sites analysed
Automatically detected principles
Web pages analysed