



Quality of health related information on the Web

Machine learning approach for automatic quality criteria detection of health web pages

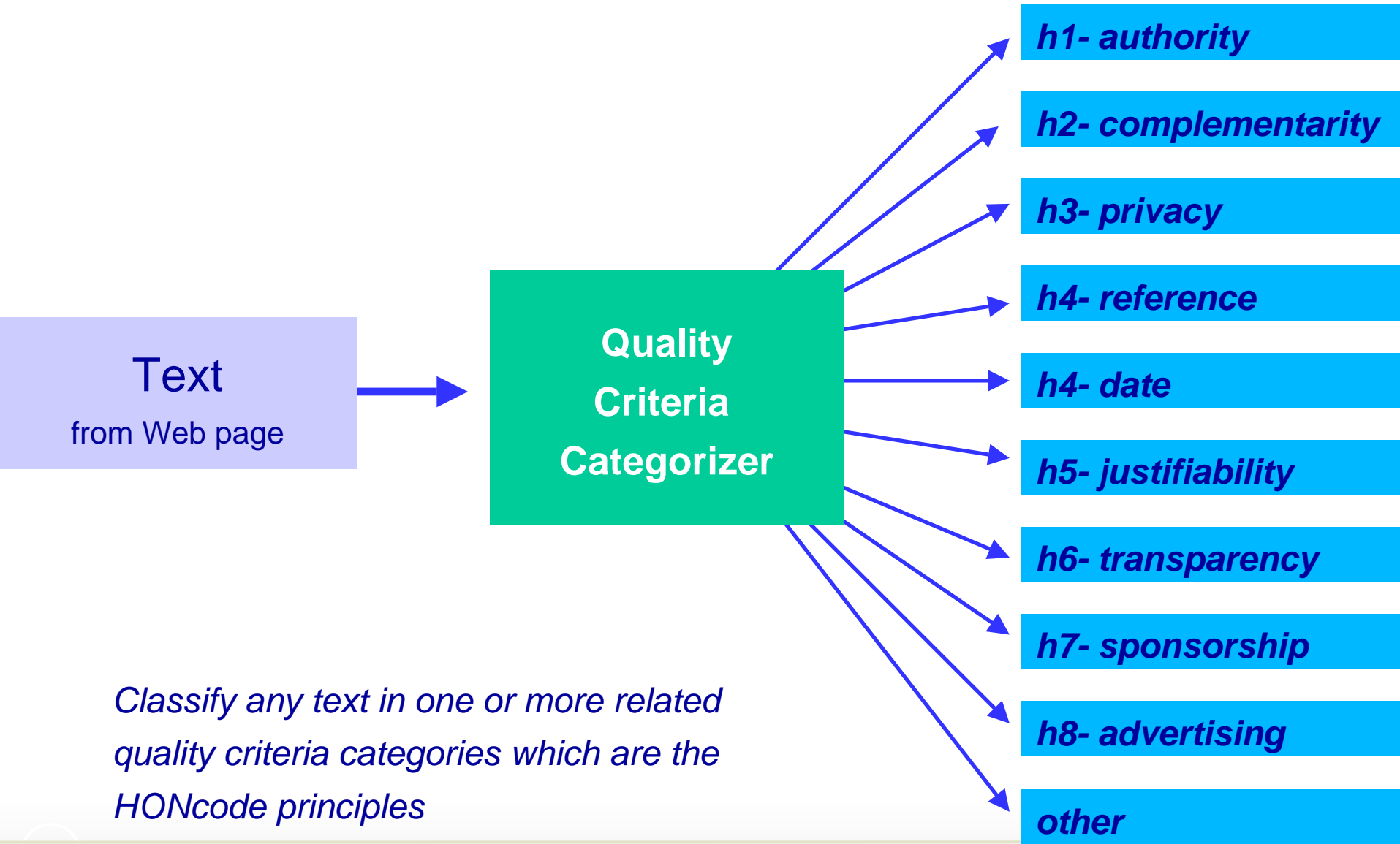
An original approach for accessing the quality of Web information

Arnaud Gaudinat, Natalia Grabar & Celia Boyer

Brisbane, MEDINFO 2007



www.hon.ch



Enter Query

Readability Filter

Trust Filter

Readability Filter: easy_to_read

Trust Filter: Authority | Privacy

Found 992 Trusted Pages

Results from Trusted Site 1 to 10

KWIC

Medicines To Prevent Asthma Attacks



... medicines need to be taken even when there are no obvious symptoms. They are particularly helpful in **preventing asthma** attacks due to allergies, exercise, cold air, and some air pollutants. By reducing the swelling and mucus... Children Last updated September 2004 Page 7 of 16 . Other helpful websites: . Medicines To **Prevent Asthma** Attacks. Medicines are taken daily, whether or not symptoms occur, to **prevent** attacks. The goal of these medications is to reduce the inflammation...

www.yourmedicalsourc.com/library/asthmachild/ASC...

Last visited: 3-1-2007

<http://www.yourmedicalsourc.com>

Trusted by Source:

[HONcode](#)

2- AAAAI - Patients Consumers Center: Tips to Remember: Prevention of Allergies and Asthma in Children



... current information and can match this current information with the needs of your family. . Other considerations in **preventing asthma**. Maternal smoking during pregnancy is associated with increased wheezing during infancy. Exposing children to secondhand smoke in... children. This pamphlet describes steps that may be taken to delay or, possibly, **prevent** the onset of allergies and **asthma** in children. **Preventing** food allergies. Food allergies in children can cause a variety of problems that...

www.aaaai.org/patients/publicedmat/tips/prevention...

Last visited: 5-1-2007

AAAAI - Patients Consumers Center: Seniors and Asthma



... Some of those can't be **prevented**, but at least you can take a flu shot every fall. Being aware of your **asthma** triggers can help control your **asthma**. Your doctor can help you explore... , is a practicing allergist in Los Alamos, NM, and a Fellow of the American Academy of Allergy, **Asthma** and Immunology (AAAAI). . . .

[www.aaaai.org/patients/seniorsandasthma/asthma tri...](http://www.aaaai.org/patients/seniorsandasthma/asthma_tri...)

Last visited: 3-1-2007

<http://www.aaaai.org>

Automatic Detected Trust Criteria: [Auth](#) [Priva](#)

[Medhunt](#)

Two main issues for medical information on the Web

Sponsored Links

Uneven quality



*Information
overwhelming*

Shark Cartilage 750mg

New Zealand Natural Product Online.
Free Shipping 120 Cap only \$22.95
www.greenhealth.co.nz

Premium Shark Liver Oil

Pure w/ Squalene from New Zealand's
Deep-sea. Mega-Boost Immune System!
cancerchoices.com

Shark Cartilage on Sale

Get it & Save Now
Free Shipping - Same Day Shipping!
b.com

Shark Cartilage

Powerful liquid extract
High bio-availability!
erSmart.com

Shark Cartilage

Full naturalife range of natural
products. Fast delivery.

www.DrugstoreChemist.com/naturalife



PIPS Context

- Part of PIPS (Personalized Information Platform for Life & Health) - a 4 years IST Project.
 - Goal: The overall objective is to develop and pilot a Health & Life Knowledge and Services Support Environment :
 - Healthcare Professionals in deliverance of personalized and prevention-focused healthcare services compliant with the patient's personal context of his/her health state, preferences and ambient conditions;
 - The public to make informed decisions concerning treatments and nutrition at any time and place according to the real-time evaluation of his/her state of health.
 - HON Part: "Trust mechanisms for e-Health Knowledgebase" from "Trust and Security Management" WP3
 - HON developments:
 - Trusted Search with Privacy functionalities (P3P)
 - Model of trust for PIPS Platform
 - Detector of quality criteria
 - Health trustworthy Question/Answering system
 - Health Literacy categorizer

Initiatives for accessing the quality of Web sites

- **Selection or referencing** (the first approach):
Yahoo!, MedlinePlus (Miller et al, 2000), CISMeF (Darmoni et al, 1999)
- **Self-regulation** (the most utopian):
IHC : eHealth Code of Ethics (Risk, 2000), TNO Health Trust : QMIC (Sheldon, 2002)
- **Accreditation of web pages** (the most accurate):
HONcode (Boyer et al., 1996), URAC (webapps.urac.org)
WMA (Bosch, 2002)
- **Popularity of web pages** (the most naturally used):
Google (Page et al., 1998)
- **Collaboration of users** (the most democratic):
Outfoxed / Lijit (www.lijit.com), Google co-op (www.google.fr/coop), Wikipedia
- **Education of users for quality evaluation** (the most complementary):
Oxford University : DISCERN, OMNI (Organising Medical Networked Information)

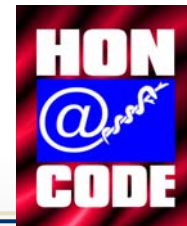
Manual accreditation at the HON Foundation

- Web site submits a request to get the HONcode seal
- Web site fills a questionnaire
- HON reviewers deeply analyse web site
 - According to finding, web site is:
 - Accredited
 - Accredited under condition to modify/add some information
 - Not accredited
- Each web site is reviewed every year
- HONcode consists of 8 principles
- HONcode base contains over 5'500 web sites

Accreditation is a extensive task => Automatic assistance needed

HONcode ethical principles

1. *Authoritative*
Indicates 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 to the site by the visitor
4. *Attribution*
Cite the source(s) of published information, date, 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, accurate email contact
7. *Sponsorship*
Identify funding sources
8. *Advertising policy*
Clearly distinguish advertising from editorial content



Automatic detection

- Automatic detection of quality of health Web sites
 - (Price & Hersh, 1999)
 - Presence of HONcode seal and other criteria
 - No evaluation and no method description
 - (Aphinyanaphongs, 2003)
 - Supervised Text categorization method on scientific literature material
 - (Griffiths & Al, 2005)
 - Original method, based on IR relevance feedback, on depression domain
 - (Wang & Liu, 2006)
 - Detection of code based on regular expressions
- Machine learning for text categorization
 - Hostile messages (Spertus et al., 1997)
 - Spam (Carreras et al., 2001)
 - Racist content (Vinot et al., 2003)

Excerpts selection by HON reviewers



Mehta childcare -- privacy policy - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

URL   http://www.mehtachildcare.com/general/privacy.htm

Mehta
Childcare

...because informed parents have healthy children

Privacy policy

I do not gather any information about you. My site offers unrestricted access to every part; there is no registration (free or otherwise). I put no cookies on your computer, and do not spy on you in any way.

If you ask a question, I reply to the email address you have used. I send one, and only one, email if you ask a question. If you ask for further information, I reply to that, too, once. No unsolicited mail, and I will never sell, distribute, or share your email addresses with anyone.

I collect the questions you ask. That is part of the payoff I get for the time invested in the site -- an insight into parents' concerns. For my personal use only.

Thanks for reading my Privacy statement, and for visiting my Website. I hope you will find it useful.

Dr. Parang N. Mehta
M.D. (Pediatrics)

Home
Asthma
Childhood diseases
The Baby Page
Vaccination
Food and drink
Childhood
Contact Dr Parang

Extract 

Terminé

Size of corpora: number of excerpts

	English	French	Spanish	Italian
<i>Authority</i>	1685	188	123	230
<i>Complementarity</i>	1738	182	119	190
<i>Privacy</i>	1561	128	106	187
<i>Attribution Ref</i>	1039	112	71	128
<i>Attribution Date</i>	2069	232	196	297
<i>Justifiability</i>	323	25	17	28
<i>Transparency</i>	1813	177	120	201
<i>Sponsorship</i>	1473	163	101	163
<i>Advertising</i>	1030	103	86	142

Features of the method

- Machine learning algorithms (*supervised text categorization*):
 - SVM, NB, kNN, DT
- Learning material:
 - Textual excerpts from accredited web sites
 - URL addresses
- Combination of the two scores: Text & URL
- Linguistic pre-processing: Stemming, Stopwords ...
- Various discrimination features: ngram, cooc, ...
- Unit of classification: Sentence of textual excerpts
- Languages processes: English, French, Spanish & Italian
- Evaluation:
 - 90% learning, 10% test
 - precision, recall, F-measure

A selection of evaluated methods and features on all classes

Lang	Lem	Meth.	Weight	maR	maP	maF1	miR	miP	miF1	Err
ENG	w1	SVM	nnn	0.65	0.74	0.69	0.69	0.78	0.73	0.06
ENG	w1	NB	nnn	0.72	0.67	0.66	0.81	0.65	0.72	0.07
ENG	cooc	NB	nnn	0.68	0.71	0.69	0.75	0.72	0.73	0.06
ENG	w1	NB	atc	0.67	0.69	0.61	0.77	0.61	0.68	0.08

Lang	Lem	Meth.	Weight	maR	maP	maF1	miR	miP	miF1	Err
FRE	w1	SVM	nnn	0.72	0.82	0.76	0.67	0.80	0.73	0.06
FRE	w1	NB	nnn	0.80	0.73	0.75	0.80	0.63	0.70	0.08
FRE	w1	DT	nnn	0.70	0.77	0.73	0.64	0.73	0.68	0.07

Precision/Recall contingency (SVM ENG W1)

Ass \ Cor	Authority	Complem.	Privacy	Reference	Justif.	Trans.	Sponsor	Advert.	Date
Authority	64 / 72	05 / 05	01 / 01	19 / 34	01 / 09	04 / 13	04 / 09	01 / 01	00 / 01
Complem.	05 / 05	80 / 82	05 / 03	01 / 02	06 / 44	00 / 00	03 / 05	00 / 01	00 / 00
Privacy	02 / 03	02 / 04	92 / 90	00 / 01	00 / 03	01 / 02	01 / 02	02 / 06	00 / 00
Reference	24 / 13	03 / 02	03 / 01	64 / 57	02 / 08	01 / 01	02 / 02	00 / 00	01 / 02
Justif.	06 / 01	32 / 03	06 / 00	06 / 01	45 / 33	02 / 01	00 / 00	00 / 00	00 / 00
Trans.	06 / 02	02 / 01	08 / 02	02 / 01	00 / 00	81 / 81	00 / 00	01 / 00	00 / 00
Sponsor.	05 / 03	04 / 02	02 / 01	01 / 01	00 / 02	02 / 02	69 / 69	16 / 17	00 / 00
Advert.	01 / 01	02 / 01	05 / 01	00 / 00	00 / 02	00 / 00	13 / 12	77 / 73	00 / 00
Date	00 / 00	01 / 00	01 / 00	06 / 03	00 / 00	00 / 00	01 / 01	01 / 01	90 / 98

- *Globally results are very good for Privacy and Attribution_ref*
- *Results are good for Complementarity, Transparency, Sponsorship and Advertising*
- *Small confusions between Sponsorship and Advertising*
- *Small confusions between Authority and Reference*
- *Confusions between Justifiability and Complementarity*

Results summary

- SVM gives better results for precision
- Classical tf-idf Weighting is better (raw frequency nnn)
- Stemming has no effect (to verify for other languages)
- Usage of selected stop-words improve results
- About principles:
 - *Privacy, Attribution_date* are very well recognized
 - *Authority, Complementarity, Privacy, Transparency, Sponsorship and Advertising* have good results
 - *Attribution_ref* and *Justifiability* have bad results
- French, Spanish and Italian have promising results
- Introduction of a medical general class is efficient

- Supervised categorization
- Categorization into nine classes
- Independent from a specific medical domain and language
- Learning database is continuously updated by specialists

- Suitability for assisting accreditation and revision process
 - Approach will soon be integrated in review process

- Integration and Relevancy in search engine should be evaluated
 - A prototype of Search Engine already exists (with quality filter facilities)

- Main difficulty is due to the difference between the discrimination unit level (sentences) and the final classification level (page or site)

- Improvement of *Justifiability* through the use of MEDLINE references
- Merge with a classical global approach based on a collection of bad and good pages



worldwide

Health information in a multicultural world

<http://www.HealthOnNet.org/>



*Thank you very much for
your attention!*

This work has been realized as part of the PIPS (personalized Information Platform for Life & Health) project funded by the European Commission program IST- 507019.

Evaluation results

- Promote precision and accuracy

Lang	Lem	method	weigth	maR	maP	maF1	miR	miP	miF1	Err
ENG	w-1	NB	xxx	0.72	0.67	0.66	0.81	0.65	0.72	0.07
ENG	w-2	NB	xxx	0.69	0.71	0.66	0.79	0.66	0.72	0.07
ENG	w-3	NB	xxx	0.71	0.65	0.65	0.79	0.62	0.69	0.08
ENG	w-4	NB	xxx	0.69	0.59	0.60	0.77	0.52	0.62	0.11
ENG	cooc	NB	xxx	0.68	0.71	0.69	0.75	0.72	0.73	0.06
ENG	s-1	NB	xxx	0.72	0.68	0.66	0.81	0.64	0.72	0.07
ENG	w-1	NB	nxx	0.71	0.71	0.65	0.80	0.64	0.71	0.07
ENG	w-1	NB	tfc	0.64	0.73	0.61	0.75	0.62	0.68	0.08
ENG	w-1	NB	xxc	0.65	0.61	0.61	0.76	0.60	0.67	0.08
ENG	w-1	NB	tfx	0.76	0.57	0.64	0.79	0.63	0.70	0.08
ENG	w-1	SVM	xxx	0.65	0.74	0.69	0.69	0.78	0.73	0.06
ENG	w-2	SVM	xxx	0.61	0.74	0.67	0.66	0.79	0.72	0.06
ENG	w-1	KNN	xxx	0.45	0.71	0.53	0.48	0.81	0.61	0.07
FRE	w-1	NB	xxx	0.80	0.73	0.75	0.80	0.63	0.70	0.08
FRE	cooc	NB	xxx	0.77	0.76	0.76	0.77	0.70	0.73	0.06
FRE	w-1	SVM	xxx	0.72	0.82	0.76	0.67	0.80	0.73	0.06
FRE	w-2	SVM	xxx	0.68	0.87	0.75	0.62	0.85	0.71	0.05
FRE	cooc	SVM	xxx	0.71	0.77	0.73	0.66	0.73	0.69	0.06
FRE	n-1	DT	xxx	0.70	0.77	0.73	0.64	0.73	0.68	0.07
SPA	n-1	NB	xxx	0.54	0.47	0.49	0.67	0.51	0.58	0.11
SPA	cooc	NB	xxx	0.50	0.50	0.49	0.62	0.58	0.60	0.09
ITA	n-1	NB	xxx	0.67	0.54	0.58	0.81	0.60	0.69	0.08
ITA	cooc	NB	xxx	0.60	0.58	0.58	0.73	0.69	0.71	0.07

Evaluation results (Precision) on 'English SVM simple word'

- Privacy and Date are well recognized (HC3 and HC4-date)
- 32% of Justifiability assigned was Complement
- 24% of Reference assigned was Authority

	Authority	Complement	Privacy	Reference	Justifiability	Authorship	Sponsorship	Advertising	Date
Authority	0.64	0.05	0.01	0.19	0.01	0.04	0.04	0.01	0.00
Complement	0.05	0.80	0.05	0.01	0.06	0.00	0.03	0.00	0.00
Privacy	0.02	0.02	0.92	0.00	0.00	0.01	0.01	0.02	0.00
Reference	0.24	0.03	0.03	0.64	0.02	0.01	0.02	0.00	0.01
Justifiability	0.06	0.32	0.06	0.06	0.45	0.02	0.00	0.02	0.00
Authorship	0.06	0.02	0.08	0.02	0.00	0.81	0.00	0.01	0.00
Sponsorship	0.05	0.04	0.02	0.01	0.00	0.02	0.69	0.16	0.00
Avertising	0.01	0.02	0.05	0.00	0.00	0.00	0.13	0.77	0.00
Date	0.00	0.01	0.01	0.06	0.00	0.00	0.01	0.01	0.90

Evaluation results (Recall) on 'English SVM simple word'

- 98% of Attribution Date is well recognize
- 90% of Privacy is well recognize
- 34% Reference is mis-recognised as Authority (HC4)
- 44% Justifiability is mis-recognised as Complement (HC5)

	Authority	Complement	Privacy	Reference	Justifiability	Authorship	Sponsorship	Advertising	Date
Authority	0.72	0.05	0.01	0.34	0.09	0.13	0.09	0.01	0.01
Complement	0.05	0.82	0.03	0.02	0.44	0.00	0.05	0.01	0.00
Privacy	0.03	0.04	0.90	0.01	0.03	0.02	0.02	0.06	0.00
Reference	0.13	0.02	0.01	0.57	0.08	0.01	0.02	0.00	0.02
Justifiability	0.01	0.03	0.00	0.01	0.33	0.01	0.00	0.00	0.00
Authorship	0.02	0.01	0.02	0.01	0.00	0.81	0.00	0.00	0.00
Sponsorship	0.03	0.02	0.01	0.01	0.02	0.02	0.69	0.17	0.00
Avertising	0.01	0.01	0.01	0.00	0.02	0.00	0.12	0.73	0.00
Date	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.01	0.98