Multi-theme automatic quality detector for health web pages

H. Lagahzi, A. Gaudinat, C. Boyer
Celia.Boyer@healthonnet.org

The context
Uneven quality of health information online: "in the realm of cancer diagnosis and management, a large portion of the information is incomplete, misleading or both" Lee CT et al, Bladder cancer facts: accuracy of the information on the Internet. J.Urol. 2003;170:1756-60

HONcode and certification process
1. Authority
2. Complementarity
3. Privacy
4. Attribution
5. Justifiability
6. Transparency
7. Financial disclosure
8. Sponsorship

Previous research in automatic detection
Wang et al.: a method of detecting indicators for quality of health information
Gaudinat et al.: Combination of heterogeneous criteria for the automatic detection of ethical principles on health websites

Rationale
• The need to detect the quality of the content, i.e. not limiting detection to how content points to a principle or its adherence to it
• To address the entire health web and not a limited number of health websites

Materials
• Topics selected from Quackwatch by a physician
• Verification of the scientific conclusion with Pubmed
• Manual annotation of the content: low or high quality corresponding to objectivity and relevance. Evaluation conducted by an MD
• HTML pages are cleaned to keep only the text content: HTML codes and advertisements have been removed
• Manual identification and categorization by health theme and quality level of 3,326 English web pages from 136 websites

Methodology
• Support Vector Machine (SVM) algorithm
• 10-fold cross validation method

Features studied
Number of words: counts the number of words in each document
Compression ratio: evaluate the content redundancy. Documents were compressed using “gzip” and the ratio of the size of documents before and after compression was calculated. Low quality sites could be more redundant
Adjectives and adverbs: evaluate the document sensitivity. We used a list of 730 adjectives and 3732 adverbs. We calculated the proportion of adverbs and adjectives on each document. Our hypothesis is that low quality sites could be more sensitive
Medical terms: we used a Mesh extractor to evaluate the proportion of medical terms in documents, where “bad” sites could have less medical terms

Discussion
• The ratio of medical terms showed a significant result
  - Among documents containing less than 40% of medical terms, low quality-category documents are represented more
  - Significant difference between “high” and “low” quality-category documents considering the medical term ratio
• Feasibility and efficiency of automatic quality detection

Limits of the study
• Web pages/sites have an exclusive correspondence with classes
• Possible bias: categorizer which classify site instead of classes
• 10 folder cross-validation evaluation may not be enough

Conclusion
• Assessment of web pages: a very challenging task
• Automatic categorization with learning approach could be further prospected
• But difficulty of identifying efficient discriminative features which are domain independent
• Next Steps: increase the number of unscientifically-based health information

The table above shows the indicators used to evaluate the model and the score associated
• The error rate is less than 2%
• This table highlights the results obtained for the “high quality” classification
• Considering precision, this table shows that the system is capable of 98%. It means that for the “high quality” classification, in 100 pages returned by the system, 98 are well classified