Multi theme automatic quality detector for health web pages

Hind Lagahzli[a,b]**, Arnaud Gaudinat**[b]**, Célia Boyer**[b]

[a] *Polytech Grenoble*
[b] *Health On the Net Foundation, Geneva, Switzerland*

Abstract and Objective

In the health information field, mechanisms to help assess quality appear to be more than ever a high priority need. The increasing volume of health information available online covers a large spectrum of health related topics. However, most of the time internet users have no indicators about the reliability of this information. State of the art approaches to address this problem consist of either machine learning algorithms or automatic evaluation with simple regular rules. In this poster, we present our multi theme reliability level based on a large corpus of low- and high quality web pages representing several health domains, complementary to our supervised detector based on generic quality criteria. We attempt to classify health web pages independently of the health domain, classifying them according to a scale of reliability. This study shows that an automatic recognition of low quality health pages was possible, with less than 2% error rate. However the question of domain dependency is still remaining and further studies are needed to evaluate the system's ability to deal with new health domain documents.

Keywords:

Trust, Health web reliability, Quality, Text categorization.

Methods

Many attempts of evaluating the quality of online health documents have been done. The manually approach is efficient but the number of evaluated websites is limited. For instance, the Health On the Net Foundation has defined the HONcode, ethical and quality code of conduct with 8 principles. The HONcode promotes quality and trustworthy health information on the web. Despite its accuracy, this approach is time limited and human resources consuming. Our study presents a machine learning approach for a quality detector system applied to health web pages. The questionable topics were selected from the Quackwatch site by a physician. For each topic, pages of high and low quality content were selected. Features such as the number of word or the medical term ration have been studied to determine whether they are discriminatory, in order to classify the quality of health documents independently of the health domain.

A support Vector Machine (SVM) algorithm was used. This algorithm was reported in previous studies, to be the most suitable one for several text categorization tasks. Different combinations of parameters, features selection and word combination (e.g. n-grams of 1 to 4 words), were experimented on to obtain the best results.

To evaluate the system performance, we used a 10 fold cross validation method. Data, previously prepared, were randomly divided to 10 different sets, each one representing 1/10 of the corpus documents. Training set contained 9/10 of data while the 1/10 remaining documents were used as a test set. Test sets were alternatively chosen from the 10 different data sets, while training set contained the 9 left sets each time.

Results

Documents containing less than 40% of medical terms are most often categorized, “bad”[i]. Documents containing more than 40% of medical terms are most often in the category "good". To evaluate the learning model performances, the following indicators were used: precision, recall and F-measure. For each indicator, we choose to represent both macro and micro values. The macro values are representative of the distribution of elements in each category, while micro values reflect the distribution in each document. The error rate is also computed. We calculated the average of the 10 cross validation tests results. We could highlight the “good” results obtained by the SVM learning machine for this corpus of data, especially the error rate inferior to 2 %.

In the first part, only the ratio of medical terms showed a significant result. We were expecting a higher compression ratio sensitivity score for “bad” category documents, as observed in Spam indexing. However, we showed a significant difference between “good” and “bad” category documents considering the medical terms ratio. In the second part, the feasibility and the efficiency of automatic quality detection have been shown on this corpus with only 2% of error.

Limitation: there is a possible bias, the creation of a categorizer able to classify websites instead of classes. Thus a 10 cross folder classical evaluation on all documents could sometimes not be sufficient, and the test folder should contain only web pages or sites never used in the learning set.

[i] Unproven treatments or information