A method for automatic content classification in health informatics based on specialized thesaurus

Fabio O. Teixeira[a], Alex J. Falcão[a], Anderson D. Hummel[a], Felipe Mancini[a], Thiago M. Costa[a], Fernando S. Sousa[a], Domingos Alves[b], Ivan T. Pisa[c]

[a] *Postgraduate program in Health Informatics, Federal University of São Paulo (UNIFESP), São Paulo - SP*
[b] *Social Medicine Department, Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto – SP*
[c] *Health Informatics Department, UNIFESP, São Paulo - SP*

Abstract and Objective

The purpose of this study is to present the results of a procedure for automatic classification of scientific articles in Health Informatics using a specific thesaurus. Statistical, vectorial, and artificial intelligence methods were applied to classify HI-related content. Statistical procedures and measures of accuracy, precision, recall, area under the ROC curve, and F1 measures were performed to measure the degree of similarity between terms of the specialized Health Informatics thesaurus and the selected articles. The percentage of accuracy achieved was 0.87, F1 measure was 0.88 and the area under the ROC curve was 0.94. The study results were positive showing a remarkable difference in the classification patterns, based on a specialized HI thesaurus, between specialized HI articles and those from Health.

Keywords:

Vocabulary, Controlled, Classification, Artificial intelligence.

Introduction

The experiment focus is to distinguish between Health and Health Informatics (HI) articles using well-know methods (1) and a specialized HI thesaurus, named Health Informatics Thesaurus (HIT) (2), for classifying articles, then, recognizing the relevance of HIT's terms to Health Informatics domain.

Methods

The Health Informatics Thesaurus (HIT) contains 730 terms, of which 620 were obtained from HI technical scientific literature. The remaining 110 terms were extracted from the *Medical Subject Headings* (MeSH).

This thesaurus can be accessed and downloaded at http://telemedicina6.unifesp.br/project/teixeif.

Titles and abstracts of 900 HI scientific papers were collected. These articles were published between 2006 and 2009 and it was distributed uniformly in 9 journals. It was also collected 900 articles from Health, it was published between 2004 and 2009 and it was distributed uniformly in 9 journals too. After, statistical, vectorial, and artificial intelligence methods were applied to the collected articles and specialized HI thesaurus (HIT) for classification of HI-related content (3). In our experiments, six classifiers were used: Artificial Neural Networks, Support Vector Machines, K-nearest neighbors, Naive Bayes, Bayes Net, and Decision Trees. To perform the experiments, the 10-fold cross validation method was adopted.

Results

The Artificial Neural Networks classifier provided the best results, where accuracy, precision, recall, AUC, and F1 measures were 0.87, 0.86, 0.89, 0.94, and 0.88, respectively, obtained as results of the applied method.

Conclusion

The authors believe that classification task of articles with HI content could be improved based on similarity with Health Informatics Thesaurus terms.

References

[1] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. 2 ed. Morgan Kaufmann; 2005.

[2] Colepicolo E, Novaes MAP, Pisa IT, Wainer J. Epistemology of the Medical Informatics [Internet]. 2005 [citado 2009 Fev 2]. Available from: http://www.icml9.org/program/poster5/

[3] Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. J. Am. Soc. Inf. Sci. Technol. 2006;57(1):96-113.