ITPP0816
ICT

ARC'14
مؤتمر مؤسسة قطر السنوي للبحوث
QATAR FOUNDATION ANNUAL RESEARCH CONFERENCE
Towards World-class Research and Innovation

Automatic Category Detection Of Islamic Content On The Internet Using Hyper Concept Keyword Extraction And Random Forest Classification

10.5339/qfarc.2014.ITPP0816

Abdelaali Hassaine; Ali Jaoua

Corresponding author :
hassaine@qu.edu.qa
Qatar University, Doha, Qatar

Abstract

The classification of Islamic content on the Internet is a very important step towards authenticity verification. Many Muslims complain that the information they get from the Internet is either inaccurate or simply wrong. With the content growing in an exponential way, its manual labeling and verification is simply an impossible task. To the extent of our knowledge, no previous work has been carried out regarding his task.

In this study, we propose a new method for automatic classification of Islamic content on the Internet. A dataset of four Islamic groups has been created containing texts from four different Islamic groups, namely: Sunni (Content representing Sunni Islam), Shia (Content representing Shia Islam), Madkhali (Content forbidding politics and warning against all scholars with different views) and Jihadi (Content promoting Jihad). We collected a dataset containing 20 different texts for each of those groups, totalizing 80 texts, out of which 56 are used for training and 24 for testing.

In order to classify those contents automatically, we first preprocessed the texts using normalization, stop words removal, stemming and segmentation into words. Then, we used the hyper-concepts method which makes it possible to represent any corpus through a relation and to decompose it into non-overlapping rectangular relations and to highlight the most representative attributes or keywords in a hierarchical way. The hyper concept keywords extracted from the training set are subsequently used as predictors (containing either 1 when the text contains the keyword and 0 otherwise). Those predictors are fed to a random forest classifier of 5000 random trees.

The number of extracted keywords varies according to the depth of the hyper concept tree, ranging from 47 keywords (depth 1) to 296 keywords (depth 15). The average classification accuracy starts at 45.79% for depth 1 and remains roughly stable at 68.33% from depth 10. This result is very interesting as there four different classes (a random predictor would therefore score around 25%).

This study is a great step towards the automatic classification of Islamic content on the Internet. The results show that the hyper concept method successfully extracts relevant keywords for each group and helps in categorizing them automatically. The method needs to be combined with some semantic method in order to reach even higher classification rates. The results of the method are also to be compared with manual classification in order to foresee the improvement one can expect as some texts might indifferently belong to more than one category.