Automatic Estimation of Human Values Using Supervised Machine Learning

学位　　博士（情報科学）
(Ph.D. in Information Science)
取得大学　九州大学 (Kyushu University)
取得年月日 March 25, 2015
高山泰博 (Takayama, Yasuhiro)

This thesis investigates issues and solutions regarding estimation of "human values" that are reflected on statements in contentious documents, aiming to apply the solutions to contents analysis in social science. Social scientists have used findings from content analysis to provide information for public opinion surveys and decision-making by governmental agencies on contentious issues. Content analysis is typically performed by trained human annotators and is conducted by means analyzing written or transcribed documents. Annotators can detect the human values that the writer expressed in textual form. For example, they analyzed the values of stakeholders who support or oppose the idea of "net neutrality" as expressed in testimonies prepared for public hearings for U.S. legislative bodies and regulatory agencies. This research was based on a subset of the meta-inventory of human values (MIHV). The human values specified in the inventory include six human values: *freedom*, *honor*, *innovation*, *justice*, *social order*, and *wealth*. Content analysis is a type of qualitative analysis, thus typically subject matter experts must conduct the analysis. Hence it is fairly costly and analysis remains restricted because analysis of a large number of documents is not feasible. Therefore, the goal of this research is to explore novel methods for automatically detecting human values invoked in textual documents using machine learning.

For facilitating content analysis, several inventories of human values are used in social science research. Integrating key components of these studies, An-Shou Cheng and Kenneth Fleischmann defined human values as follows: "values serve as guiding principles of what people consider important in life." And they developed MIHV, which is intended to be applicable to various test collections by selecting values specific to the debate at issue and by iteratively refining annotation guidelines.

Social scientists considered that human values were reflected on statements rather than words, thus traditional paper-based annotations for values could annotated any length of passages. In addition annotated passages often overlap. Cheng et al. constrained each annotation to be a single sentence but allow for more than one value per sentence. This set up is well suited to supervised machine learning. They applied MIHV and sentence-based annotation to the net neutrality corpus. After several round of refining the annotation guideline, the resulting 9,890 sentences in the original corpus, containing 102 documents, are annotated with zero or more of six human values. We adopt this net neutrality corpus as our test collection. For making the data more tractable, we remove longer sentences and sentences whose boundaries are uncertain, then stemmed them. Finally, we obtained the remaining 8,660 sentences.

Emi Ishita et al. have reported the classification of human values for the earlier round of the net neutrality corpus using *k*-NN (Nearest Neighbor) classifiers for 2,005 sentences over 28 documents. They obtained a macro-averaged F_1 (a harmonic mean of precision and recall) of 0.48 for eight human values. To scale up social science research, we need to improve the ability to effectively analyze larger data corpora. The lexical features must serve as a basis for classification of human values. We first compared the effectiveness of a wide range of classifiers available within the machine learning tool Weka, and we found that Support Vector Machines (SVMs) performed best for our corpus.

The human values estimation resembles sentiment analysis, which has been extensively researched. An important difference is that estimation of human values is a multi-label classification, whereas sentiment analysis is typically modeled as a binary classification. Importantly, human values can help to explain

sentiment, given their explanatory power in relation to attitudes and behavior. We focus on human values estimation problem with the net neutrality corpus and our modeling of relation between sentence-level labels and word-level human values in this thesis.

In our corpus, the number of occurrences of two-thirds of distinct words is less than five, and the average number of words per sentence is only 10.3. That is, we face two kinds of data sparseness for estimating human values in the net neutrality corpus: (a) the number of sentences is small for training classifiers i.e. eight thousand sentences or so; (b) the number of words per sentence is very small, i.e. ten or so within a sentence. We have to accept the above (a) because our purpose is to substitute human annotators with machine classifiers to reduce the costs of annotation. However, it is challenging to estimate the relationships between sentence-level values and the words within the sentence for detecting values of the sentence, because most of the words occurred rarely. In addition, the conventional best performed SVMs are general purpose binary classifiers, thus it is unclear whether SVMs can appropriately deal with relationships between the multiple values annotated in a sentence and words as constituents of the sentence. Therefore, estimation of values remains a matter of research to explore classifiers with higher accuracy which is specialized for detecting values.

The contributions of this thesis for estimating human values are as follows. Firstly, we found that augmenting feature vectors which include hypernyms and synonyms did not substantially contribute to improving classification effectiveness of values in experiments using SVMs with augmented feature vectors. For detecting sentence-level values, it is revealed that we could not deal with training data sparsity by augmenting features using word semantic categories, in spite of the fact that word semantic categories are effective for word sense disambiguation or syntactic disambiguation. The fact suggested us the direction of a new method that assigns the values directly to words, instead of assigning the values to semantic categories.

Secondly, we formulated the human values for a sentence as an aggregation of human values for words that are constituents of the sentence. Then we proposed a probabilistic latent "value" model (LVM), that automatically switches the case where estimation of human values for a word is influenced by the previous word and the case where the estimation is done without influence from the previous word. We achieved that approximately three percent relative improvement in F_1 score by the proposed method, compared with one by the conventional method that used an SVM with simple bag-of-words features, and we confirmed that the improvement is statistically significant. We also demonstrated that the classification effectiveness of LVM is comparable to the second human annotator. This means that human annotators might be replaced by classifiers that use our proposed LVM.

Thirdly, we built up a "values-dictionary" which consist of words and word pairs with human values for the use by social scientists, and we demonstrated the possibility of support them to conduct content analysis. The classification effectiveness of our proposed LVM achieved high scores, however, it is difficult for humans to interpret the estimation results because the estimation using LVM is done by integrating probabilities of words (or word pairs) representing the values in that sentence. Therefore, we proposed a simplified model of LVM, where the model extracts words (or word pairs) with human values as entries of a values-dictionary, that are uniquely determined independently from a specific sentence. The model is an approximate probabilistic optimization whose objective function is F_1 score for detecting sentence-level values. In addition, we have provided the values-dictionary to social scientists for content analysis, and we convinced that the values-dictionary is usable for a tool of extracting actual annotation examples that would be suitable for the annotation guideline.