Mining and Organizing User-Generated Content
to Identify Attributes and Attribute Levels

by
Artem Timoshenko and John R. Hauser
MIT Sloan School of Management
77 Massachusetts Avenue
Cambridge, MA 02139
(atimoshe@mit.edu, hauser@mit.edu)

Conjoint-analysis applications are only as good as the attributes and attribute levels upon which they are based. In the best applications, conjoint-analysis attributes are relevant to managerial decisions and relevant to consumer choice. For example, practitioners might rely on an extensive study of consumers that includes in-depth qualitative interviews. When pre-studies are done well, practitioners identify attributes that are relevant to both managers and consumers. However, these steps can take time and can be costly. We explore a faster and less-costly alternative.

Increasingly, consumers (and experts) discuss products online. This user-generated content (UGC) includes reviews, blogs, tweets, Facebook posts, expert recommendations, and forums. For most products, the corpus of online content is reasonably comprehensive—it can be mined for attributes and attribute levels. However, the sheer size of the corpus presents problems—typically, tens of thousands (or even more) discussions are readily available for each product category. Analyzing a human-readable sample of the corpus may not sufficiently represent all attributes and attribute levels. But mining the corpus is not enough. We must winnow and organize the corpus.

We demonstrate an automated machine-learning approach to mine, winnow, and organize UGC. Step 1. We use a convolutional neural network (CNN) to classify sentences in the UGC corpus as “informative” vs. “non-informative.” The CNN is “trained” on a small sample of sentences classified by a human coder, but thereafter is automated. Step 2. We winnow the set of informative sentences using an available set of pre-trained word vectors. These pre-trained word vectors assign a numerical score to every word to capture semantic meaning. For example, \( v(\text{queen}) \approx v(\text{king}) - v(\text{male}) + v(\text{female}) \). For each informative sentence, we use the word vectors to form a sentence vector. By clustering the sentence vectors, we identify groups of semantically-similar sentences. Each cluster becomes a customer need. Step 3. We organize the
customer needs into a hierarchical structure of primary, secondary, and tertiary customer needs. The structure is based on semantic similarity and is automated. Conjoint-analysis attributes and levels are then drawn from the hierarchy subject to the focus of the conjoin-analysis study. (This last step is manual so that the conjoint analysis is focused on the appropriate managerial question.)

The automated methods identify a hierarchical structure of customer needs, but are they relevant? We seek to examine the veracity and appropriateness of the hierarchy. We do so by comparing the UGC-based customer needs and hierarchy to customer needs and a hierarchy obtained by tradition voice-of-the customer (VOC) studies. The comparison VOC study is drawn from an industry application and is based on (a) focused experiential customer interviews, (c) transcripts highlighted by trained analysts, (c) winnowing by group processes, (d) sorting of needs by customers, and (e) hierarchical clustering of a co-occurrence matrix.

50-word Description

Using machine-learning methods applied to online user-generated content, we identify, winnow, and organize informative content to produce a hierarchical structure of customer needs. The structured needs identify attributes and attribute levels that are relevant to managerial decisions and consumer choice.