Extracting Meaning from Cell Phone Improvement Ideas

Jenine Turner
Athenahealth
jturner@athenahealth.com

Raimondas Lencevicius
Qwobl
raimondas@acm.org

Mark Adler
Nokia Research Center
mark.adler@nokia.com

Abstract

Companies have recently begun gathering product improvement ideas via web tools. Resulting data collections are too large to be effectively dealt with by human users. But natural language processing and machine learning techniques are well suited for this type of problem. We explore several ways to organize such data in the cell phone domain: supervised classification, unsupervised clustering, and time-based analysis.

Numerous companies nowadays gather product improvement ideas. Reviewing all of the resulting thousands of ideas without tools would require a great deal of time and resources. Automatic tools can help these reviewers in a number of ways. The questions we address here are categorization, finding common ideas, and finding idea trends over time. We explore techniques to answer these questions using suggestions from the cell phone domain. Each idea is presented to us as a title along with free text.

Semi-supervised categorization (Routing)

A large organization will have multiple groups working on different aspects of a product, and we would like to use NLP tools to route ideas to their appropriate product group. This would then be a simple classification task except that we do not have labeled training data. Nor, in fact, did we start with well-defined categories. Our solution was to turn to cell phone manuals, which provide helpful words in various categories.

Categories and Training Data

To gather our categories, we looked at four cell phone manuals, and chose categories that appeared frequently across the tables of contents. There were 14 categories in all: Applications, Battery, Connectivity, Contacts, Hardware, Log, Maps, Media, Messaging, Organizer, Settings, UI, Voice, and Web. We also have a *none* category, mainly for ideas that are junk or gibberish, but also for ideas that have no words overlapping with other ideas or training data.

We used the cell phone manuals to gather training data as well, by gathering the text for each category from the relevant section.

Classification method

The features we use are unigrams and bigrams from the text (treated as separate features, without interpolation).

Once we have created feature vectors, our classification method is straightforward. For each idea vector that we are categorizing, we choose the category of the training vector that is most similar, according to the cosine distance:

$$cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

There are two additional modifications we use to adjust our feature set, that provide improvements over the original feature counts. The first is based upon our assumption that words in the title are more important than words in the other text fields. We simply weight unigrams and bigrams that appear in the title ten times as heavily as those that appear in the rest of the text.

The other modification of the feature set that we employed was to weight words according to their *selectional preference strength* (Resnik 1996). Selectional preference of a word is defined as:

$$S(w) = D(P(C|w)||P(C)) = \sum_{c} P(c|w) log \frac{P(c|w)}{P(c)} \quad (2)$$

C is the category, and we assume uniform probabilities over categories. The selectional preference is the KL-divergence of $P(C|w)$ and $P(C)$, which just tells us how different these distributions are. If a word is evenly distributed (*phone*, for example), then the probability of the category given the word is the probability of the category. If the word is *camera*, then the probability of the media category given *camera* is much higher than the probability of media (and the probability of all the other categories given *camera* is much lower than the probability of the category alone). Thus words that strongly prefer certain categories have a high selectional preference strength. By multiplying the counts in the word vectors by the words' selectional preference strength, we emphasize important words.

Evaluation

We hand-annotated 100 ideas with zero to three labels. There was one idea labeled with zero labels (which is categorized as *none*), 77 with one label, 20 with two labels and

one with three labels. Our best results came from using both the title weighting and the selectional preference weight (.75 accuracy). Weighting the title alone gives .67, using selectional preferences alone gives .71, and cosine distance alone gives .63.

Unsupervised clustering

Labeled classification is useful, but since users can enter free text on any subject, ideas may not solidly fall into any specified category. Unsupervised clustering methods yield themselves well to discovering structure that may not be as obvious in classification. We use k-means and agglomerative clustering, as well as sub-clustering our semi-supervised categories.

When performing Group Average Agglomerative Clustering (Manning & Schuetze 1999), instead of the normal starting point of having each idea in its own cluster, we use the results of k-means clustering as starting points. We find that starting with each idea in its own cluster causes some initial mistakes to be compounded in future merges.

Another use of unsupervised clustering is to subdivide the categories that result from our semi-supervised classification. We use k-means to cluster each category, allowing one new subcluster for every 25 ideas. This makes the resulting clusters more specific than the results of regular k-means, but this specificity is what we are looking for.

Labeling and Tagging

Since large numbers of clusters can be hard to explore, we discuss labeling schemes for the clusters created by the unsupervised clustering methods mentioned in the previous section. We also discuss a technique for tagging individual ideas with the most pertinent words, which is useful as a summary of the idea.

Labeling clusters The goal of labeling the clusters is to give more structure to unsupervised clusters which would otherwise be more difficult to navigate. A good label is a word that appears frequently in one cluster, but rarely elsewhere. In k-means clustering we compare a cluster against all other clusters, and in agglomerative clustering we compare a cluster to its sibling cluster, which has the effect of emphasizing the differences between two clusters that are joined.

(Connection, WLAN, network), (Play music, track, listening), (Maps, GPS, data, location, position), (Maps, route, nagivation, car, GPS), and (Color, blue, theme, black, white) are some examples of labels assigned to our unsupervised clusters.

Tagging ideas Whereas labels provide a summary of the cluster, tags provide a summary of a single idea. We generate a keyword list - relevant words for our ideas data, by gathering frequent labels over 20 runs of k-means. Some examples of keywords are *picture*, *charging*, *video*, *sms*, *contacts*, and *browser*. We also use the fact that the title tends to be more relevant, and if a word on the keyword list appears in the title, it becomes a tag. In addition, we use words that are generally infrequent but appear multiple times in the idea.

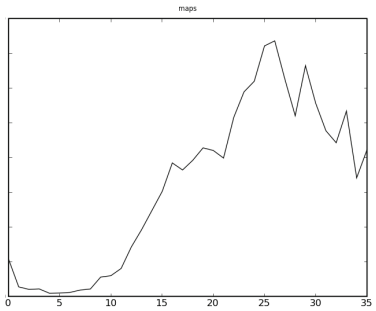

Figure 1: *maps* over time

Time-based analysis

A question of particular interest to those choosing product direction is what sorts of trends exist in the data. Have some kinds of ideas become more popular recently and should therefore become a focus? Are there other ideas that are no longer coming up, indicating that perhaps an issue was solved, and resources devoted to that subject can be directed elsewhere?

One way to get at the notion of trends is to find out what is popular in a given month. Unusual words might indicate trends. We thus created lists for each month's most indicative words. Some of these words include *itunes*, *carkit*, *voip*, *firefox*, and *podcast*.

To get a high-level sense of the changing of frequency of topics over time, we also create a smoothed, normalized graph of our keywords. Many keywords show no particular trend towards becoming more or less common over time, but others show increasing or decreasing popularity, or a spike at a particular time period. Figure 1 gives an example of the keyword *maps*, increasing in popularity over time.

Discussion and Future Work

We have shown good empirical results in semi-supervised classification based on cell phone manuals. Unsupervised clustering allowed us to find patterns not obvious from the semi-supervised classification. We were also able to approach trend analysis by finding month-specific words, and looking at trends overall in frequent keywords. Our future interests include discovering better trends, and modeling cluster movement over time. For this purpose, we are also interested in gathering more data.

References

Manning, C., and Schuetze, H. 1999. *Foundations of Natural Language Processing*. Massachusetts Institute of Technology.

Resnik, P. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* 61:127–159.