Mining Web Text for Brand Associations

Robert Malouf
Department of Linguistics and Oriental Languages
San Diego State University
5500 Campanile Dr
San Diego, CA 92182-7727

Bradley Davidson, Ashli Sherman
CommonHealth
Wayne Interchange Plaza II
155 Route 46 West
Wayne, NJ 07470

Abstract

Weblogs and other web text are an incredibly rich knowledge base, and the marketing industry is beginning to recognize the value of web texts as a source of information about their customers. However, the nature of web texts make them unsuitable for analysis using standard market research methods. In this paper, we describe the use of exploratory data analysis techniques to extract the associations patients and their caregivers have for eight leading medications for a seizure disorder. We first collect a list of candidate keywords which occur with a brand name or a variant spelling, and then cluster the keywords to construct a set of brand associations. To compare individual brands we measure the association between each brand name and any term from each of the clusters. Finally, using dimensionality reduction techniques, we plot the brand names, their associations, and their relationships as a brand association map.

Web text, such as blogs, newsgroups, message boards, and email lists, can provide an easily collected and incredibly rich source of data on a nearly limitless range of topics. Issues related to health and medicine are particularly well represented: as of October 25, 2005, Yahoo! Groups (`groups.yahoo.com`) listed 85,349 separate discussion groups under the "Health and Wellness" category, and Technorati (`www.technorati.com`) lists 79,367 blog posts with the tags "Health", "Health and wellness," or "Medical". People faced with medical problems turn to the Internet for support from other patients and their families, for information about their disease and the possible treatments, for help navigating with the medical establishment, and sometimes for a sympathetic venue to vent their frustrations.

Not surprisingly, the marketing industry is beginning to recognize the value of web texts as a source of information about their customers. However, the nature of web texts make them unsuitable for analysis using standard market research techniques. The sheer quantity of data makes comprehensive qualitative analysis impossible. While looking at a small sample of text may yield some insights, that may also lead an analyst to grant to much weight to accidental properties of a small sample, and at the same time to miss subtle patterns which can only be detected by looking at the data set as a whole. Web texts contain a huge amount of information in the aggregate, but individual posts by themselves are rarely informative.

Computational linguists and other researchers in the field of text mining have developed a toolkit of statistical methods for analyzing large quantities of text. In this paper, we describe the use of a combination of rule-based and statistical exploratory data analysis techniques to extract brand associations: the feelings, beliefs, knowledge, and attitudes people have towards a product. The focus of the study was to find and document the associations that patients and their caregivers have with eight medications for a seizure disorder.

The corpus we are working with is a collection of posts to a number of Internet discussion groups and other websites used by epilepsy patients and their families. The corpus contains a total of 26,062,526 words in 316,373 posts from 19 different sites and 8,731 distinct users. Posts average 119 words each.

"Sentiment analysis" is a well-known technique for measure whether the associations with a brand name are generally positive or negative (Turney 2002). This is unlikely to be helpful for looking at brand associations with medication names, however. One reason for this is that the associations with medication names, especially those used to treat a serious condition like epilepsy, are overwhelmingly negative. Even if someone would recommend a treatment to another patient, on balance almost all patients would much prefer not to be taking any medications at all. Another problem with applying sentiment analysis in this domain is that patients turn to on-line communities when they have a problem. Users rarely post to say that their disease is still under control or that they are suffering no noticeable side effects. So, the overall direction of the sentiments expressed in web texts will be strongly biased toward the negative, simply by the nature of the medium. And, finally, a simple negative or positive judgment is too coarse-grained. Market researchers need to know more detail about brand associations, and in particular are interested in what differentiates between competing brands in the minds of consumers.

What is needed in this application to make use of web texts for market research is a set of quantitative techniques for extracting brand association patterns from large quantities of unstructured web text. The methods need to be sufficiently automatic that they can be applied to very large quantities of data without human intervention. They also must be flexible enough to allow room for manual intervention by a domain expert when appropriate. Finally, the results need to

be visualized in way which makes them comprehensible to someone who is not a text mining expert, and which makes clear what actions should be taken based on the evidence in the web texts.

The first step in extracting brand associations is to identify posts which mention one of the target brand names. As a first pass, we could simply perform a keyword search, retrieving all posts which contain one of the brand names. FDA regulations require that medication names not be easily confused with normal words or with each other, which means that this strategy will result in few false positive matches. However, medications are often marketed under a variety of different names. For example, the anti-convulsant Tegretol is known by the generic name "carbamazepine" and is also has been sold under the brand names Amizepine, Carbazepin, Epitol, Finlepsin, and Neurotol. To expand the search to include these alternate names, we extracted the alternate entries for each medication from MeSH, a controlled vocabulary for subject headings developed by the US National Institute of Health's National Library of Medicine (`http://www.nlm.nih.gov/mesh`).

A more serious problem with a naive keyword matching strategy for finding brand name mentions is that users' spellings of these words vary widely. Web texts in general are informal and unedited, and are full of idiosyncratic spelling, punctuation, and formatting. In addition, medication names are by nature unfamiliar, difficult to spell words, making the likelihood that they will be spelled correctly in web texts even lower. And, finally, posters often use conventional nicknames for commonly mentioned drugs rather than their full official name.

To test the effectiveness of various keyword spotting strategies, we constructed a high-recall list of candidate brand name mentions. Using a finite-state transducer, we extracted all terms from the corpus which were with a Levenshtein distance of 3 or less from one of the MeSH names for carbamazepine (e.g., *tegretol*, *tegreatol*), or which is within a Levenshtein of one or less of being a prefix or a suffix (e.g., *teg*, *tege*). This produced a list of 644,129 candidate mentions, with 516 unique types. We then filtered this list by hand to find 6,150 genuine brand mentions (with 97 types).

Using this hand-corrected list as a gold standard, we can estimate the precision and recall of a variety of keyword spotting techniques. The naive algorithm, simply searching for mentions of "Tegretol," yields 4,492 hits, for a recall of 73.0%. Adding the alternative names extracted from MeSH increases this to 4,728 hits, for a recall of 76.9%. Adding words with an Levenshtein distance from one of the names of one or less raises recall to 86.0%, with a negligible decrease in precision. Adding words which are a prefix of "Tegretol" further increases recall to 91.7%, still with near perfect precision. Moving to a Levenshtein distance of two further increases the recall, to 96.7%, but the precision drops to a worrying 90.8%. Therefore, in the following analysis, we considered to contain a mention of a brand name all posts with a which is word within an Levenshtein distance of one from, or which is a prefix of, one of the names listed in MeSH.

The next step is to collect a set of candidate *keywords*. These are words which (potentially) reflect the issues surrounding the brand names which users find salient. We proceed by marking all posts which contain a mention of at least one brand name, and compute the pointwise mutual information (Church & Hanks 1989) with each term.

After calculating this PMI score for each of the 20,505 vocabulary items which occurred at least fifteen times and which are not themselves brand names, we selected the top 5% as potential keywords. This yields 1,001 terms which co-occur with brand name mentions much more frequently than would be expected strictly due to chance.

Given a list of candidate keywords, we next construct a set of keyword clusters which reflect the issues potentially associated with brand names in the original corpus. As a first step, we construct a term co-occurrence matrix listing the number of times each word in the corpus occurs within a 15-word window of a 'content bearing word'. For our purposes, a content bearing word is a non-function word which occurs in the corpus with moderately high frequency.

Next, we reduce the dimensionality of the term co-occurence matrix using Latent Semantic Analysis, a statistical technique similar to Principle Components Analysis, to reduce the influence of random noise in the data and to extract distributional patterns among words which reflect their semantic relationships (Schütze 1997). In this reduced dimensionality WORDSPACE model, each term is represented as a vector of 100 latent variables which reflect the distribution of terms in the original co-occurence matrix.

We next perform a complete-linkage hierarchical clustering of the candidate keywords based on their WORDSPACE representations, at each stage joining a keyword or cluster of keywords with its closest neighbor in the semantic space. We define the semantic distance between two words as the cosine of the angle between their WORDSPACE vectors. The distance between two clusters is the longest pairwise distance between any two members of the clusters.

For our list of 1,001 candidate keywords, hierarchical clustering yields a set of 161 keyword clusters containing 999 keywords (keywords which do not fit into any cluster are dropped). Manual inspection of these clusters reveals that many of them represent plausible brand associations. For example, we find a MEMORY cluster that reflects the cognitive side effects of many anti-convulsant medications:

> *loss memory problem cognitive term short concentration speech trouble confusion recall concentrate coordination inability*

and a DEFECTS cluster reflecting possible risks for pregnant women:

> *pregnancy pregnant risk birth defects women pregnancies risks baby childbearing dangerous trimester fetus*

Other keyword clusters are semantically plausible, but the connection to the brands is not obvious. An example is the QUICKLY cluster:

> *finally eventually quickly fast awhile*

And, a few clusters, such as the PROV cluster, are hard to account for:

fitly wbschool prov apples steven spoken silver settings brandy gold

The keyword clusters found in the previous step represent topics of discussion that come up in the corpus of web texts in connection with the brand names. To compare individual brands, we can now compute the PMI for the association between individual brand names and any term from each of the clusters. The top ten clusters for some of the brands under investigation are as follows (some of the brand names have been altered to protect proprietary information):

- Tegretol
  CARBATROL DEFECTS DRUGS BLOOD PARTIAL SEIZURES YRS KG CARERS CP
- Zonakote
  EPILESPY MG RASH DEFECTS DOSE CARBATROL SWITCHING KATHY NIGHT JME
- Lamantin
  WEIGHT DRINKING PROV PAIN EFFECT EAT MEMORY DOSE HEADACHE SHAWN

Some of these associations can be easily explained. For example, CARBATROL is a medication that contains carbamazepine, the active ingredient in Tegretol. Zonakote can cause insomnia and vivid dreams (NIGHT) and is connected to Stevens-Johnson syndrome, a potentially fatal side effect marked by a distinctive rash (RASH). Lamantin is also prescribed for migraine headaches (MIGRAINE) and eating disorders (EAT), and its side effects include memory problems (MEMORY), weight loss (WEIGHT, EAT), and dehydration (DRINKING).

Other associations are more interesting. All medications in this class carry some risk for pregnant women. However, in the web texts, Tegretol and Zonakote are more closely associated with discussion of these risks (DEFECTS) than Lamantin is. Similarly, finding the right medication for a patient and then finding the right dose requires a process of trial and error for all anti-convulsants. Yet, Zonakote is more strongly associated with dosage related issues (MG, DOSE, SWITCHING) than Lamantin is.

And, as is probably inevitable for an unsupervised text mining method, some of these associations are specific to the web texts which happened to be selected for the corpus. The keywords in the SHAWN cluster (*shawn emily multiple tle hemiplegic*) are connected to individual participants who happen to be taking Lamantin. This does not reflect a general pattern, but may still help provide some insight into what is going on the the web texts. For a conventional unsupervised semantic learning task, text clusters like these would reflect noise in the data. For a web mining task, however, these clusters represent something about the structure of the texts themselves. An analyst interested in learning more about how users talk about Lamantin may want to look more closely at posts mentioning words in this cluster. The process of figuring out why these words tend to associate with Lamantin in these texts may yield new insights into the general patterns. If nothing else, this will help focus qualitative analysis on a small subset of the posts in the corpus (out of a total of 316,373 posts, only 970 mention Shawn).

Effective visualizations are vital part of any exploratory data analysis technique. A simple way to visualize the associations found in the previous section is to present the PMI scores as a series of bar graphs, one for each keyword cluster. This allows the domain expert to identify at a glance which clusters show a skewed association towards one or more brand names and are worth further investigation.

Latent semantic analysis (as used in section 2) has been used as the basis for effective visualizations in other applications (Widdows, Cederberg, & Dorow 2002). The 100 dimensional WORDSPACE representation of the brand names and the keyword cluster centroids can be plotted in a two dimensional plane using multidimensional scaling, producing a plot which shows the general semantic relationships among the terms. This technique however does not highlight the kind of brand associations that we are interested in. The brand names under investigation all turn out to be much more closely related to each other than they are to any of the keyword clusters. LSA highlights the similarities between the terms, but we are primarily interested in their differences.

Rather than base our visualizations on the 100 dimensional WORDSPACE model, we instead construct a vector for each of the brand names based on the PMI scores with the keyword clusters. This allows us to interpret the combinations of associations for each brands as a location in a 160 dimensional 'association space', which we then reduce to two dimensions using Independent Component Analysis (Marchini, Heaton, & Ripley 2004).

The methods described in this paper have been used to investigate the brand associations that patients and their caregivers have for a number of medications based on a large collection of web texts. These methods are used both to further explore results gained from more conventional market research and to generate new hypotheses that can then be verified by other methods. In future work, we will apply these techniques to other market research tasks, such as tracking brands over time, monitoring the impact of advertising or other marketing initiatives, and for improving customer resource management strategies.

References

Church, K. W., and Hanks, P. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27 Annual Meeting of the ACL*, 76–83.

Marchini, J. L.; Heaton, C.; and Ripley, B. D. 2004. *fastICA: FastICA algorithms to perform ICA and Projection Pursuit*. R package version 1.1-6.

Schütze, H. 1997. *Ambiguity Resolution in Language Learning*. Stanford: CSLI Publications.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424.

Widdows, D.; Cederberg, S.; and Dorow, B. 2002. Visualisation techniques for analysing meaning. In *Fifth International Conference on Text, Speech and Dialogue*, 107–115.