Using Latent Semantic Analysis to Identify Successful Bloggers

Aron Price
Ohio State University
Cognitive Systems Engineering
price.561@osu.edu

Manirupa Das
Ohio State University
Supercomputer Center
das.65@osu.edu

Annatala Wolf
Ohio State University
Computer Science Engineering
wolf.332@osu.edu

Abstract

Recently, research has been done on identifying the influential bloggers in a community based on links between blogs. It would be interesting to identify whether characteristics of informal written text alone, such as vocabulary or word choice, could be used to predict how popular a weblog will be. We cast this as a semantic problem and attempt to use latent semantic analysis to predict the comment density of comments left by both registered and unregistered users. Initially, we cluster the document vectors in the reduced space using k-means++, but this approach produces results that suggest the resulting document clusterings were topic-based. We then create a classification system to predict the comment density of unseen blogs by constructing simple vector space models of high-density and low-density blogs. The initial results of the classification approach are above chance levels, and suggest future direction for this line of research.

1 Introduction

Recently there has been much research interest in Web 2.0 technology, such as personal communication through weblogs (blogs) (Macdonald et al., 2008). Many blog posts and news articles enhance the ability of users to express opinions by allowing public comments. It is reported that 22 percent of Internet users post comments to online newsgroups or blog sites (Pew Internet & American Life Project, 2008). These comments can be viewed as an indication of the interest users have in a blog site, since they reveal the number of visitors interested enough in the blog content to post a response.

There has been recent interest in identifying the influential bloggers in a weblog community by analyzing the blog network (Agarwal et al., 2008). However, it would be interesting to be able to identify successful informal texts in data where this information is not available, such as email messages, or public blogs that do not allow comments. We hypothesize that there may exist characteristics of language use by informal writers, such as vocabulary or word choice, that are directly associated with successful communication.

Specifically, we hypothesize a relationship between the vocabulary of a blog and comment density. In this study, we use latent semantic analysis (LSA) to reduce the dimensionality of a term-document matrix for each blog in a collection (where a blog is a concatenated set of blog entries). We then perform two separate experiments. First, we use an unsupervised clustering approach to see if relationships to comment density naturally emerge from this analysis. The results of this approach suggest that naive clustering attempts end up clustering documents by topic and subtopic rather than communication style. Second, we attempt a supervised classification method to identify high and low comment density blogs, by using two complimentary models built through LSA. The results of this approach are above chance levels, and suggest potential future directions for this research.

This paper is organized as follows. In Section 2, we discuss the problem formulation and corpus used. Section 3 describes the clustering experiment and results. Section 4 describes the classification experiment and results. Section 5 discusses related work. We conclude in Section 6 by summarizing contributions and discussing future directions.

2 Problem Formulation

Information retrieval on blog systems can be difficult: blogs are filled with neologisms, opinion information, and informal speech (Macdonald et al., 2008). It would be interesting to identify features in blog text that are correlated with the success or *popularity* of a blog, because this information could be used to refine the prior probability of a blog's importance in probablistic models of information retrieval. Information on word choice and popularity would also be relevant in sociology for developing models of successful communication styles.

Many blogs contain public information that identifies whether the blog is popular, such as rating systems, recommendations, or comment counts. Other blogs can be judged to be influential by examining the structure of inbound links to the blog (Agarwal et al., 2008). However, not all blogs contain this information in a form where it is readily available.

We focus on the problem of identifying non-sequential features of blog text that correlate with the *success* of the blog, as measured by the mean number of comments per entry, or *comment density*, generated by people who read the blog. We consider the comments of registered users and unregistered users separately. Our assumption is that high comment density, especially when resulting from a large number of different registered users, indicates a blog is more popular.

2.1 Blog Corpus

The corpus we use is drawn from Wordpress[1], which allows us access to a large number of blogs with spam screened out. Each blog is from a unique Wordpress domain. Our tests were performed on a single template group (the Mistylook theme). Non-English blogs were removed. LingPipe[2] was used for the tests.

3 Clustering Tests

When used on a large number of documents, an unsupervised approach with LSA is often useful for clustering documents by topic (Landauer et al., 1998). The initial experiment attempted to determine whether LSA could be used to identify characteristics of vocabulary associated with successful bloggers. The presence of multiple topics would be a potential confound for this association, as LSA would likely cluster based on words that occur in the most popular topics.

To help ensure that an unsupervised approach would not cluster by topic or subtopic, the set of blogs were restricted to a narrow focus. First, three categories were hand-selected from a list of 30 most common keywords in wordpress: politics, religion, and movies. These categories were selected because they seemed to represent a distinct set of topics. For each category, the LSA tool at CU Boulder[3] was used to identify the twenty nearest-neighbors in English. These words were counted as topic markers.

Blogs were excluded from consideration if they contained fewer than 11 entries, to prevent new or abandoned blogs from introducing noise into the category analysis. Remaining blogs were then categorized into topics if they met a threshold, d, of minimum term density (category term count divided by total word count). This threshold was raised until there was little overlap between categories. This first attempt only left a small number of blogs, and all were in the politics category. Since politics was the most common category identified, the movies and religion categories were removed from consideration entirely. The threshold d was relaxed to 0.0003, which produced a sample size of 760 blogs.

A term-document matrix was generated, and LSA was run twice: once with 500 dimensions, and once with 10 dimensions. The purpose of the 10-dimensional reduction was to test whether a severe reduction would perfom a more reliable clustering. Clustering in a high-dimensional vector space can produce results which vary widely, so seeding is an important step. K-means++ is a method which improves on the seeding of traditional k-means, which improves the probability of selecting robust clusters (Arthur and Vassilvitskii, 2007).

3.1 Results

Clustering was performed with two, three, and five clusters, both on the 10-dimensional and 500-dimensional document vectors. K-means++ was performed five times for each test, and a voting method was used to determine the outcome. In all cases, a majority of the five tests produced identical clusters.

[1] http://wordpress.com/

[2] http://alias-i.com/lingpipe/

[3] http://lsa.colorado.edu/

The difference between the cosine similarity scores of the average vectors among the two clusters was not very large. Additionally, analysis of the clusters showed clearly that the clusters were largely topical in nature: in all cases, one of the two clusters was focused almost exclusively on religion, even though the topic had been restricted to blogs containing political terms. This indicated that the blog corpus was not restricted enough to a single topic in order to allow for LSA clustering to isolate non-topical information, though this may not have been an effective approach even in a highly constrained topic space.

4 Classification Tests

For the second experiment, we employed a supervised classification approach. Blogs with more than 5 entries and non-English blogs were removed from consideration. The remaining 3,277 blogs were separated into training and testing in a standard 70/30 proportion.

For each blog in the training set, total comment density was calculated. The median value of comment density among the training set blogs was used as a cutoff, and blogs were separated into high-density and low-density training categories (of 1,147 blogs each). For each category, LSA was run at 300 dimensions, and a term-document matrix was produced. All of the document vectors in a category were summed together to produce a model vector for that category.

Each blog in the test set was tested against both models. To test a blog, each term appearing in the blog entries was checked against the term matrix for one of the models. If the term appeared in the matrix, it was added to a sum vector for the blog/category test. This sum vector was then compared with the model vector for that category using cosine similarity. The category model that produced the highest similarity score was the category predicted. If the prediction was high-density, this was counted as a hit when the total comment density exceeded the training data cutoff, and vice versa.

4.1 Results

The results are summarized in Table 1. Overall classification accuracy was 57.7%, which is significantly above chance level. Most of the misses came from mistaken misprediction of low-density for high-density results.

Table 1: Classification Results

Density	Hits	Misses	Total	Percent
Low	374	124	498	75.1
High	193	292	485	39.8
Total	567	416	983	57.7

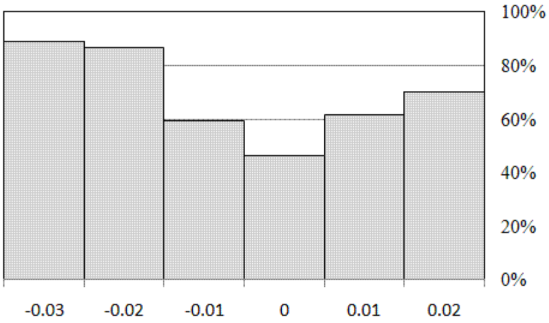

Figure 1: % Hits by Cosine Difference

The plot in Figure 1 illustrates how prediction accuracy improves with distance from the median cutoff value.

4.2 Discussion

Classification accuracy seems promising, but there is no basis for concluding that LSA is not simply selecting topics of high and low popularity. Determining whether topicality can be removed from consideration would require additional tests, such as the restriction of features to non-content-bearing words (such as stopwords). Even with this kind of approach it would be difficult to show that topic had no bearing on the features being selected by the system.

However, if this result can be improved upon, it may be a useful technique for information retrieval even if topic selection is the primary source of prediction. It may prove useful to attempt the same test using bigrams, if a large enough training set can be gathered to produce a sizeable set of bigrams.

Low-density prediction accuracy was nearly twice as high as high-density prediction accuracy. This suggests that blogs which have few comments may be more semantically distinct than blogs with many comments.

5 Related Work

Recently, there has been much interest in identifying characteristics of bloggers which may act as indicators of the value of the contained in their blogs. Recent articles that have directly addressed the influence of blogs have done so by examining the link structure that exists between blog entries (Agarwal et al., 2008). However, not all blogs feature inlinks and outlinks, particularly in a blog community whose focus is less social than expert. Our study seeks to identify features correlated with blog popularity without considering the links between blogs.

Other studies have addressed the expertise and credibility of bloggers (Balog et al., 2008; Weerkamp and de Rijke, 2008). Our work differs from these in that we consider the entire text of the blog as a feature for classification. Additionally, our approach is topic independent: we don't care if the bloggers who are influential discuss multiple topics in the same blog. These bloggers may still be watched and commented on by a large number of users.

This study also shares some similarity to studies that attempt to classify the kind of blog entries, such as research that attempts to identify narrative stories (Gordon and Swanson, 2009). Our research is predominantly interested in whether blogs which become popular are classifiable through similar machine learning methods.

6 Conclusions

We have presented evidence that an LSA performed on unigram counts can be used to classify blogs as popular, as determined by a simple measurement of comment density. Our initial clustering tests were unsuccessful, and revealed how readily LSA clusters topic-based information. In retrospect, the clustering tests were not likely to be defensible even if the clusters were both separable by a wide margin and not apparently related by topic. The classification tests were successful, and low-density blogs were surprisingly much more easily identifiable than high-density blogs.

For future work, we may try building several models along a scale of varying densities, and attempt to predict the density itself rather than making a straight binary prediction. We may examine whether inclusion of bigram features improves the overall model. We also intend to examine the effectiveness of performing a similar LSA on non-content-bearing words, to see if the variable of topic can be at least partially removed from the prediction task.

References

N. Agarwal, H. Liu, L. Tang, and P.S. Yu. 2008. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218. ACM.

D. Arthur and S. Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, page 1035. Society for Industrial and Applied Mathematics.

K. Balog, M. de Rijke, and W. Weerkamp. 2008. Bloggers as experts. In *SIGIR08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*.

A.S. Gordon and R. Swanson. 2009. Identifying Personal Stories in Millions of Weblog Entries.

T.K. Landauer, P.W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Craig Macdonald, Iadh Ounis, and Ian Soboroff. 2008. Overview of the trec-2007 blog track. In *Proceedings of Text Retrieval Conference (TREC-2007)*.

Pew Internet & American Life Project. 2008. Internet Activities. [Online]. http://www.pewinternet.org/.

W. Weerkamp and M. de Rijke. 2008. Credibility improves topical blog post retrieval. *ACL-08: HLT*, pages 923–931.