Relevance Based Sorting of Forum Responses

Yash Joshi, Nandish Kotadia, Kapil Patwa, Khushali Deulkar
D. J. Sanghvi College of Engineering
Mumbai
India
yashjoshi444@yahoo.co.in, nandishkotadia@gmail.com, kapilpatwa93@gmail.com, hushali.deulkar@gmail.com

ABSTRACT: *Forums are structured environments that facilitate discussion around shared interests such as technology, sports, pet ownership, cooking techniques, etc. There are numerous forums for nearly every niche topic. Forums consist of threads, where each thread consists of an original post and responses to it. Subsets of original posts are queries related to the forum's topic. The responses to these posts are mostly expert answers due to the niche nature of the forum. However, not all responses will be answers. Users can write anything in their response post.*

Therefore, the aim of our project is separate and sort the responses based on whether it answers the original post or not. Hence we will study the different methods that can be used for this purpose and use the best suitable method.

Keywords: Naïve Bayes, Machine Learning, Text Categorization

Received: 18 December 2014, Revised 3 February 2015, Accepted 9 February 2015

1. Introduction

A growing number of websites feature a Q&A style format where questions and answers are crowd sourced. StackOverflow and MathOverflow cater to a specific technical audience while sites such as Quora and Yahoo! Answers allow questions from any discipline. These sites share similar features such as commenting and feedback systems, user rankings based upon response quality , and the ability to up-vote correct answers.

The drawback of user ranking is that the answer which are up voted most tend to be viewed by the user and up voted again hence this system builds a recursive trend of same responses to be viewed and up voted. So even if the answers that are good in content but not up voted are ignored by the users.

Consider the structure of a forum. Forum consists of threads, where each thread contains an original post followed by response posts. A subset of original posts will be questions related to that forum's topic. For example, a forum for shopping enthusiasts may contain questions about products and best deals. However, responses to forum questions are not required to answer that question. Forum users are free to respond to a question in any manner they see fit.

Q: What is the best Diwali deal on a 40" LED TV?

This user is looking for a Diwali sale on a big-screen TV. He has asked a frugally-minded community for help.Let's look at some of the responses:

R1: I'm also looking for a good deal.

R2: I am in the same market.

R3: 60k Sony at Croma is the best deal.

R4: I haven't seen any deals yet, guys.

R5: Flipkart has LED samsung 120HZ for 55k.

R6: phillips for 65k at Snapdeal.

The first, second and fourth users have shared that they are also interested in such a deal. Their responses may provide the original poster with a sense of camaraderie, but they do not answer his question. The third, fifth and sixth responses provide direct answers to the original question.

In this project we propose a system which classifies the forum responses based on their relevance to the original post.

The methodology we propose is we obtain training and testing data by implementing a web crawler and a web parser targeted against the online discussion forum and implement a classifying algorithm and machine learning technique to classify the response based on relevance.

2. Review of Literature

Naïve Bayes Classifier

Naïve Bayes[2] classifier is a simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. This algorithm computes the posterior probability of the response belongs to different classes and it assigns response to the class with the highest posterior probability. This probability model would be independent feature model so that the present of one feature does not affect other features in classification tasks.

Bayesian theory works as a framework for making decision under uncertainty - a probabilistic approach to inference and is particularly suited when the dimensionality of the inputs data is high.

Bayes theorized that the probability of future events could be calculated by determining their earlier frequency. Bayes theorem states that:

$$P\,(Y = y_i | X = x_k) = \frac{P\,(Y = y_i)\,|\,P\,(X = x_k | \,Y = y_i)}{P\,(X = x_k)} \qquad (1)$$

where:

$P\,(Y = y_i)$ - Prior probability of hypothesis Y- Prior

$P\,(X = x_k)$ - Prior probability of training data X-Evidence

$P\,(X = x_k|\,Y = y_i)$ - Probability of X given Y- Likelihood

$P\,(Y = y_i|X = x_k)$ - Probability of Y given X- Posterior probability.

Support Vector Machine

A Support Vector Machine [3] is a supervised classification algorithm that has been extensively and successfully used for text classification task. High dimensional input space: When learning text classifiers, one has to deal with large number of features. Since SVM use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.

3. Method

3.1 Algorithm

Step 1: Start.

Step 2: Obtain large data set for training purpose using web crawler.

Step 3: Apply parser on responses to obtain keywords.

Step 4: Assign labels 0 or 1 based on whether the response is answer or non-answer.

Step 5: Obtain frequency of keywords in responses which are answers.

Step 6: Apply Classification algorithm on new responses.

Step 7: Sort the forum based on relevance to the original post.

Step 8: Repeat step 6 – 7 for all new responses.

Step 9: Stop

We will crawl threads whose original posts were questions about product sales, similar to the question above (refer introduction). We manually assigned a labeling of 1 or 0 to each response depending upon whether or not that response answered the original question. We used our manually-assigned labels to train a Naive Bayes classifier.

Let $R_1 \ldots\ldots R_N$ be the list of responses to all questions. Each R_i is tokenized into a set of white-space delimited strings, and each string undergoes a series of preprocessing steps:

1. Strip all non-ASCII characters

2. Map numeric prices (e.g. \$100, \$0.50,\$499, etc.) to a single common term, PRICE TERM

3. Stem words (e.g. “*prices*”→ “*price*”→ “*reviewing*”→ “*review*”→“ *lowered* ”→“*lower*”,etc.)

4. Add the concatenation of adjacent terms to the terms list (e.g. “*no deal*” forms three terms “*no*”, “*deal*”, and “*nodeal*”)

The Naive Bayes prediction is given by:

$$p\,(R_{N+1} = 1|\,S) \tag{2}$$

$$= \frac{p\,(S|R_{N+1} = 1)\,.\,p\,(R_{N+1} = 1)}{p\,(S)} \tag{3}$$

$$= \frac{1}{1 + \dfrac{\prod_i^M p\,(S_i|R_{N+1} = 0)\,p\,(R_{N+1} = 0)}{\prod_i^M p\,(S_i|R_{N+1} = 1)\,p\,(R_{N+1} = 1)}} \tag{4}$$

The frequencies of each term in the non-answer set of responses and the answer set of responses are used to train the classifiers. For a new response R_{N+1}, consisting of preprocessed strings $S = s_1, s_1 \ldots\ldots\ldots s_m$.

4. Results and Discussions

We initially removed common words from the feature set, but found that the classifiers accuracies improved when they were left in (4.8% improvement). Concatenating adjacent words also led to a 7.5% improvement, as described in the methods section.

This may be accounted for by the observation that many English words are negated by the word immediately before them (e.g. “no deal”). We performed leave-one-out cross validation on the 8 questions. The overall accuracy demonstrated with the 254 responses was 81% (206/254 correctly classified responses) for the Naive Bayes classifier. The Naïve Bayes classifier correctly classified 76% (58/76) of the answers and 83% of the non-answers (148/178). The difference between the answer and non-answer classification accuracies was likely due to the large number of non-answer responses relative to answers.

Although the classifier accuracies were not exceptional, the utility of the algorithm was clearly visible by sorting the responses by their classification probabilities. For the example given in the Methods section, the algorithm produced the following sorting:

R3: 60k Sony at Croma is the best deal.

R6: phillips for 65k at Snapdeal.

R5: Flipkart has LED samsung 120HZ for 55k.

R1: I'm also looking for a good deal.

R4: I haven't seen any deals yet, guys.

R2: I am in the same market.

We found that particular terms were more indicative of an answer classification over a non-answer classification. These terms included "*deal*", "*PRICE TERM*", and (interestingly) "*at*". The latter of these may have contributed to the 4.8% boost that was found when common words were left in the algorithm.

5. Conclusion & Future Scope

Online forums provide a valuable resource for asking questions about niche topics. Our project demonstrates a preliminary method for organizing responses to forum questions based upon whether the response answers the original question. Since we restricted our consideration to questions pertaining to a particular category, a natural extension of this project is to apply the algorithm to arbitrary questions. This will require taking into consideration the content of the question itself; a problem in the domain of natural language processing. We speculate that this will require a separate clustering algorithm in order to group questions by related semantics. Responses to clustered questions (Such as all questions beginning with "When" or "Where") could be used to train a classifier.

References

[1] Abajian, Aaron. (2013). Sorting Forum Responses by Relevance to Original Post.

[2] Hassan, Sundus., Rafi, Muhammad., Shaikh, mMuhammad Shahid.,Comparing. (2007). SVM and Naïve Bayes Classifiers for Text Categorization with Wikitology as knowledge.

[3] Morariu, D., Cretulescu, R., Vintan, L. (2010). Improving a SVM Meta-classifier for Text Documents by using Naïve Bayes, *Int. J. of Computers*, *Communications & Control*, 5.

[4] Pratiksha, Y., Pawar., Gawande, S. H. (2012). Member, IACSIT, A Comparative Study on Different Types of Approaches to Text Categorization.