Text Normalization and Classification System for Internet Forum

University of Computer Studies, Yangon
Ya Min Htet, Thi Thi Soe Nyunt
yaminhtethlaing@gmail.com

Abstract

Internet forum is one of the most common modes of knowledge sharing through text. An internet forum is an online discussion site. From a technical point of view, forums are web applications managing user-generated text contents. Text normalization is converting 'informally inputted' text into the canonical form, by eliminating 'noises' in the text and detecting paragraph and sentence boundaries in the text and take case restoration and suggest valid words for each invalid word in the text by using dictionary. Text classification is the process of grouping text item into related predefined classes or categories to make it easier for the user to find it. The system intends to normalize and classify the internet forum. For text normalization, Cascaded approach is used and for classification Naïve Bayes (NB) method is used. In the system, hold out method is used to evaluate the system's performance.

1. Introduction

Most internet forums require registration to post. Registered users of the site are referred to as members. Members are allowed to submit or send messages through the web application. Forums are composed of thread. A thread is a collection of posts, usually displayed by default from oldest to latest. A post is a collection of messages enclose into a block containing the user's details and date and time it was submitted. Members are usually allowed to edit or delete their own posts.

Forum text can be very noisy. Specifically, it contains header, footer. It may contain extra line breaks, extra spaces, question mark, full stop, and miss used full stop and question mark. It may contain words badly cased and word misspelled.

In this system, users can import forums from www.theforumsite.com , or compose forums by themselves for input the system. And then, forum's text is normalized and classified. In normalization process, users can get their forum text in canonical form and can replace their invalid words with valid word. In classification process, users can select threshold level to classify their forum. Finally, the system gives the most suitable category to forum users for their input forum.

2. Related Work

Data normalization is an important area in data mining. In data normalization, there are two kinds of data to classified, text and tabular data.

In text normalization, there are many applications like web page data normalization, etc. eClean 2000 is a tool that can normalize emails [3]. Jie Tang, Hang Li, Yanbo Cao, Zhaohui Tang research for email data cleaning with Cascaded Approach. Yi and Liu [1] define banner ads, navigational guides, and decoration pictures as web page noises.

Text classification has many applications in natural language processing tasks such as E-mail filtering, prediction of user preferences. K. Gee and T. Fawcett showed approach to filtering emails involve the deployment of data mining techniques [5][6]. Y. Yang compared a cross-experiment between 14 classification methods, including decision tree, Naive Bayes, Neural Network, Linear square fit are the best performers[8].

3. Theory Background

3.1 Text Normalization

Text mining systems are usually designed for processing texts in canonical form. Text normalization is used to get canonical form; paragraphs are separated by line breaks, sentences have correct ending with full stop or question mark, the first word in the sentence is capitalized, and all the words are correctly cased and spelled. There are many methods in text normalization like A Unified Tagging Approach, Unsupervised Text Normalization Approach, Document Centered Approach, Hindi Approach and Cascaded Approach.

3.1.1 Cascaded Approach

Cascaded approach is one of the simple and clear methods in text normalization. Cascaded approach is divided into two processes:

(1) Non-text block filtering

(2) Text normalization.

(1) Non-text block filtering

Non –text block filtering has includes two processes. They are

- **Header detection.** In header detection perform identifying and removing forum header. Forum header includes forum title, posted user name, thread number, and posted date.
- **Footer detection.** In footer detection perform identifying and removing forum footer like web links.

(2) Text Normalization

In text normalization, there are three steps involved. These are

- **Paragraph Normalization.** In paragraph normalization, non-text block filtered text is used. In those text, may include line break. Each line break is needed to check paragraph ending. If a line break is paragraph ending, remove it.
- **Sentence Normalization.** In sentence normalization, paragraph normalized text is used. It detects missing period detection and extra spaces detection.
- **Word Normalization.** This process conducts case restoration on badly cased words and correct misspelled words [5].

3.2 Text Classification

Text classification is the task of assigning documents in natural language into one or more predefined classes [7]. There are many text classification methods such as Naïve Bayes, Bayesian Network, Support Vector Machines, Neural Network and Decision Tree.

3.2.1 Naïve Bayes Classification

Naïve Bayes is a simple probabilistic classifier based on applying Bayes' Theorem with strong (naive) independent assumptions. A more descriptive term for the underlying probability model would be "independent feature model". It is not only fast and accurate but also very efficient. It is also a language independence and easy to implement classifier. It composed of two phases: training and testing.

(1) Training Phase

There are two kinds of learning: supervised and unsupervised learning. In this system, supervised learning, train with known class label items, is used. Training is important in classification. Because the quality of training set decides the performance of classification. Before training the text, removing non-text and stop word like (a, an, the) must be performed. And then, extract keywords from train items. After this phase, keywords for each category are obtained.

(2) Testing Phase

In testing phase, calculate the weight of the input forum for each category to get the most suitable category. To calculate the weight of the forum, considering a set of internet forum F belonging a set of known category C, the most probable classification of a new forum instance is obtained combining the predictions of all hypothesis weighted by their posterior probabilities. It can be obtained by

$$P(C_i | F) = \frac{P(C_i).P(F | C_i)}{P(F)} \qquad (1)$$

Where:

- *P(C_i | F)*: the probability that a given forum belongs to *i*th category.

- *P(F):* the probability of a document that is a constant divider to every calculation and we can ignore it .

- *P(C_i)*: the probability of *i*th category.

- *P(F| C_i)*: the probability of forum given *i*th category.

Forum can be modeled as sets of words, thus the P*(F| C_i)* can be calculated by multiplying the probabilities of each individual word w_j appearing in the category (w_j being the *j*th of *l* words in the forum).

$$P(F | C_i) = \prod_{1 \le j \le l} P(w_j | C_i) \qquad (2)$$

So:

$$P(C_i | F) = P(C_i) \prod_{1 \le j \le l} P(w_j | C_i) \qquad (3)$$

$P(w_j | C_i)$can be calculate as follows:

$$P(w_j | C_i) = \frac{n_c + np}{n + m} \qquad (4)$$

Where:

- p is the prior estimate of the probability
- m is the number of attribute
- n_c is the number of instances with attribute w_j and category C_i.

- n is the total number of training instances with category C_i.

The maximum weighted category $P(C_i \mid F)$ is predicted as the most suitable category.

4. Flow of the system

There are many different forum categories. In this system, General, Art & Literature, Business & Money, Computer & Internet, Entertainment, Health & Fitness, Sports are defined. 743 forums are downloaded from www.theforumsite.com for testing and training. Among them, 500 forums are used as training set and 243 forums are used as testing set. And six threshold levels (greater than equal 10, 11, 12, 13, 14 and 15) are defined. In the system, 1362 keywords are obtained from training phase. The system is implemented by two phases: training and testing. The first phase of the system, training is carried out with the following steps.

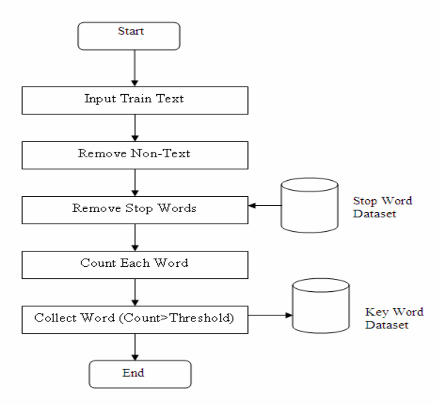


Figure 2.Flow of training phase

After training, testing phase can be started. The input of the system is a forum. The users can import the forum from testing set or compose forum by themselves. If the forum is composed, non-text block filtering step is not need to perform. Then the following steps are taken:

(1) Non-text Block Filtering. It detects the header and footer in the forum text. It then eliminates the identified blocks. Furthermore, it detects non-text like html tags and removes it.

(2) Paragraph Normalization. It identified whether or not each line break is a paragraph ending depend on the features. They are

- Whether or not the current line contains words like "Hi", "dear", "thank you", "thanks", "best regards", "sincerely".
- Whether or not the current line ends with symbols like colon, semicolon.
- Whether or not the next line starts with bullet or number or words like "thanks", "best regards", "thank you", "sincerely".

If the line break is not paragraph ending with those features, remove it.

(3) Sentence Normalization. It identifies misused question mark and full stop depend on whether or not the current sentence starts with question words by using question words dataset. If there is no space after an identified sentence ending, it adds a space there. It also removes redundant symbols (including space, question mark, full stop, exclamation mark). As a result, each paragraph is segmented into sentences.

(4) Word Normalization. It conducts case restoration on badly cased words. The starting character of a sentence must be upper case [2]. And it gives suggestion with valid words for each incorrect word by using WordNet dictionary. As a result, text in canonical form is obtained.

(5) Naïve Bayes Classification. In this step, users can choose keyword threshold level. Depends on users' choice level, calculate the weight of input forum. The most weighted category is the result or the most suitable category for the input forum. Sometimes the maximum weight is equal for more than one category. In this case, users can choose their preferences. If users satisfied the system suggested category, the forum is stored in the database.

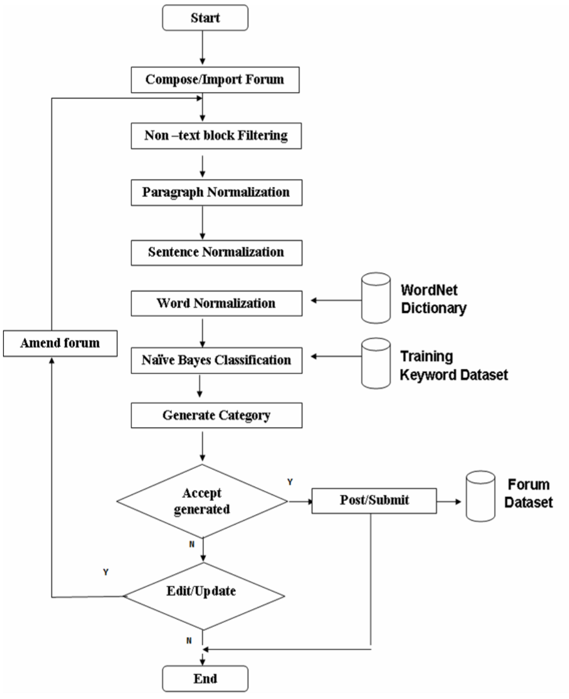


Figure 2.Flow of testing phase

5. Evaluation Methods

Performance evaluation method is important in that it allows one to evaluate how accurately a given classification will label future data, that is, data on which the classifier has no trained. Accuracy estimates also help in the comparison of different classifiers.

5.1 Holdout method

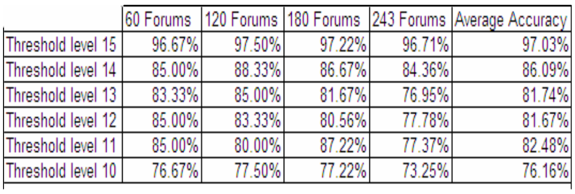
The performance of the system is evaluated by using the test result of testing data, 243 downloaded forums.

$$sensitivity = \frac{t\_pos}{pos} \qquad (5)$$

$$specificity = \frac{t\_neg}{neg} \qquad (6)$$

Where sensitivity is the probability of a positive test for one category, t_pos is the number of true positives, pos is the number of positive simples, specificity is the probability of negative test for all other categories. t_neg is the number of true negatives, neg is the number of negative samples. Accuracy is a function of sensitivity and specificity.

$$accuracy = sensitivity \cdot \frac{pos}{t\_pos} + specifictity \cdot \frac{neg}{t\_neg} \qquad (7)$$



	60 Forums	120 Forums	180 Forums	243 Forums	Average Accuracy
Threshold level 15	96.67%	97.50%	97.22%	96.71%	97.03%
Threshold level 14	85.00%	88.33%	86.67%	84.36%	86.09%
Threshold level 13	83.33%	85.00%	81.67%	76.95%	81.74%
Threshold level 12	85.00%	83.33%	80.56%	77.78%	81.67%
Threshold level 11	85.00%	80.00%	87.22%	77.37%	82.48%
Threshold level 10	76.67%	77.50%	77.22%	73.25%	76.16%

Figure 3.Testing result of the system

By the result, threshold level 15 is the most accurate level in the system.

6. Conclusion

In the proposed system, a Cascaded Approach in text normalization and Naïve Bayes method in text classification are used. A Cascaded Approach is used to clean noisy data in internet forums. Non-text block filtering is first performed on the noisy data to filter out non-text items in the data. Text normalization is then performed on the filtered data to provide cleaned data. Naïve Bayes method is used to give a suggestion that is the most suitable category for the user's forums. In the system, the accuracy rate is depends on the threshold level but not depend on the number of testing data set. The highest threshold level 15 is the most accurate level. By the system, the forum user can get their forum in canonical form and easier and efficient access to achieve required information from the internet forum.

7. References

[1] E. Xun, C. Huang, and M. Zhou. *"A Unified Statistical Model for the Identification of English baseNP"*, In Proc. of the 38th Linguistics (ACL), Hong Kong, 3- 6 October 2000.

[2] Final POESIA Workshop, *"Present and Future of Open-Source Content-Based Web Filtering"*, Pisa, IT, 21-22 January, 2004.

[3] J. Decker. eClean 2000, http://www.jd-software.com/eClean2000.

[4] Jie Tang, HangLi, Yunbo Cao, Zhaohui Tang, *"Email Data Cleaning*", Microsoft Research Asia 5F Sigma Center, China.

[5] K. Gee, "*Using latent semantic indexing to filter spam*", In Proc. of eighteenth ACM Symposium on Applied Computing, Data Mining Track, 2003.

[6] T. Fawcett, "*in vivo spam filtering, A challenge problem for data mining*", In Proc. of ninth KDD Explorations vol. 5 no. 2, 2003.

[7] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E., *"A Bayesian Approach to Filtering Junk E-Mail In learning for Text Categorization: Papers from the 1998 Workshop"*, AAAI Technical Report WS-98-05, 1998.

[8] Y. Yang, "*An Evaluation of Statistical Approach to Text Categorization*", Journal of Information Retrieval, Vol 1, No. 1 / 2, 1999, pp. 67-88.