The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

A method of sentiment analysis for online reviews containing values of multi-criteria evaluation

Takaya Nishikawa[1], Makoto Okada[2], and Kiyota Hashimoto[3]

[123]Osaka Prefecture University, 1-1 Gakuen-cho, Naka Ward, Sakai, Osaka 599-8531, Japan
(Tel: +81-72-252-1161, Fax: +81-72-254-9944)

[1]ss301018@edu.osakafu-u.ac.jp, [2]okada@mi.s.osakafu-u.ac.jp, [3]hash@lc.osakafu-u.ac.jp

Abstract: Online reviews of commercial sites are important sources for customers to obtain information and opinions. However, generally, these reviews contains several mixed information such as purpose and sentiments of reviewers. Therefore users need the method to separate these data and extract appropriate information from the reviews. In this paper, we investigated effectiveness of a method of machine learning *Support Vector Machine (SVM)* whether the method can classify reviews appropriately or not by using reviews of a travel information web site "*TripAdvisor*". We also investigate difference of precisions according to source textual data such as morpheme, bigram and trigram.

Keywords: Online reviews, Sentiment analysis, Support vector machine.

1 INTRODUCTION

The total size of the documents on the Internet has been more and more gigantic. Among them, many researchers and commercial enterprises have been paying more attention to online customer reviews found on various portal and purchase sites like Amazon.com and kakaku.com. These reviews contain a lot of personal opinions and they are considered to be very useful for other customers as well as enterprises. When a person wants to buy something, he can use others' reviews to know the product better from the customers' viewpoints. Enterprises can know better how their products are considered by customers by analysing those reviews. However, what is known with customers' reviews depends on the analytic method to be employed.

There have been many researches to analyze and utilize the reviews. Among them are two typical approaches. One is to search and extract all and only necessary information from the total data. Another approach is to classify reviews into several categories and to know the possible types of customers. For example, in some researches, weblog data are classified into two categories: *positive opinion* or *negative opinion*. In others, newspaper articles are classified into some categories based on various topics found in them.

In this paper, we adopt a different approach that classifies reviews considering not only sentiments and opinions of reviewers but also their personal information. This approach is obviously important because companies often obtain opinions classified based on generations and/or family structures of customers for market research. Therefore, it is very important for a better classification to investigate reviewers themselves together with what they write.

In this paper, we investigate effectiveness of a method to classify documents of reviews of a web site called *TripAdvisor* by experiments. The reviews there contains not only opinions but also other pieces of information such as the situation, the purpose of reviewers and themes for evaluation. For example, a keyword of "*with Friends*" or "*Solo*" are added to some reviews by reviewers. We constructed classifiers of *Support Vector Machine* (*SVM*). We constructed several classifiers by using different training data sets. We then investigate the effectiveness and the differences of each classifier based on different data sets by experimental results of review classification.

In section 2, we explain related works of our research. Next we will explain our data sets. In section 4, we explain techniques for our research, SVM. In section 5, we will show experimental results and discussion. Section 6 is the conclusion.

2 RELATED WORKS

There are two approaches to analyze a large scale document data in order to obtain useful information to human. One is a method to extract opinions, and another is a method to classify opinions.

The former approach considers components of opinions, defines the component and extracts them from text data. Sometimes the relationships of the components are used for the extraction of the opinions. Iida et al. [1] defined that an opinion can be represented as a tuple (Subject, Attribute, Value) and they extracted the pair of Attribute-value from text data.

On the other hand, the latter approach is to classify documents into some classes according to their characteristics. Hashimoto et al. [2] classified newspaper articles according to their topics or other viewpoints. In particular, *sentiment analysis* tries to classify textual data either as positive or as negative. For example, Ikeda et al. [3] classified weblogs into two classes of positive/negative.

Our research belongs to the latter approach and analyzed and classified the reviews that include much subjective information of reviewers such as feelings, reputations, and opinions of themselves. The subjective information depends on personal situations of reviewers. Therefore, in order to obtain the subjective information in the reviews considering personal data of reviewers appropriately, it is very important to classify the reviews firstly into some categories based on features or personal information of reviewers, for example single, married, have a family, and so on.

Therefore, we devised a method to classify reviews considering their personal information.

3 TripAdvisor

"*TripAdvisor*" is a review site that collects useful travel information around the world. This site contains information of hotels and restaurants in the world. And users of these facilities can add reviews to the web pages in the site. This site has more than 50 million reviews written by users in more than 20 different languages. For example, some foreign users of Japanese hotels and/or restaurants can write a review by his native language and add them to the relevant page of "*TripAdvisor*".

When users write reviews, they can add optional keywords called "Tags" that express categories of reviews. The tags show situations or purposes when writers used these hotels or restaurants. "*TripAdvisor*" provides five kinds of tags: "*Business*", "*Couples*", "*Family*", "*Friends*" and "*Solo*".

Users can also evaluate various themes by five grades. For example of the themes are "*Price*", "*Place*", "*Cleanliness*", "*Services*", and so on.

Fig.1 shows an example of reviews in this site. That shows a review text, tag data, and six themes. In later sections, we use 10,000 reviews automatically extracted from *TripAdvisor* as our data set.

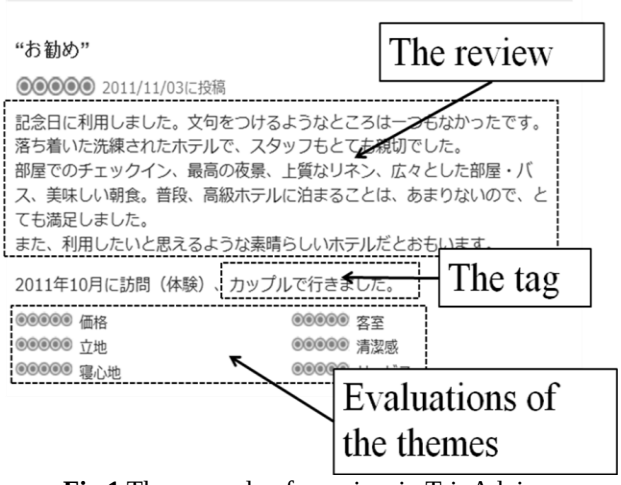

Fig.1. The example of a review in TripAdvisor

4 PROPOSED TECHNIQUES

There are various methods of machine learning such as decision tree neural network and Support Vector Machine (SVM). In this study, we selected to use SVM because it fit for our purpose. We want to classify two classes: match or not match, and SVM classify data into binary.

4.1 Support Vector Machine

A support vector machine (SVM) is a supervised learning method for binary classification, originally proposed by Vapnik [4]. Fig.2 shows the basic concept of SVM.

The algorithm computes the hyperplane which divides two classes by analyzing the vector data of two classes as teacher data. With the hyperplane, the machine estimates the class of unknown data. Even if vector is high dimensional, the machine can analyze. Since the dimension of vector of documents data is often high, the machine is used for analyzing of natural language.

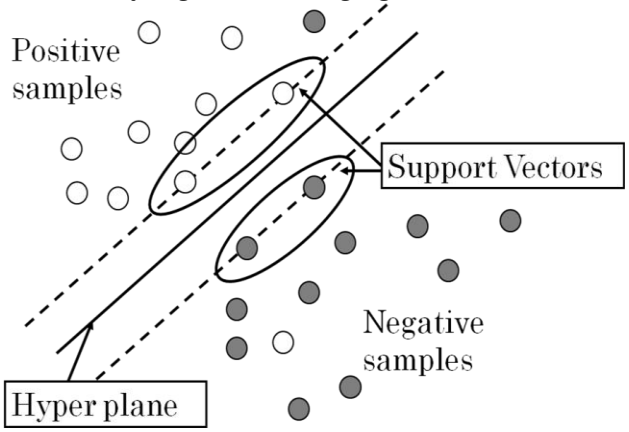

Fig.2 The illustration of SVM

4.2 Vectorization

We used ideas called "bag-of-words". For example, here are two texts S1 and S2:

S1: John likes to watch movies. Mary likes too.

S2: John also likes to watch football games.

Based on these two text documents, a dictionary is constructed as:

dictionary = {1:"John", 2:"likes", 3:"to", 4:"watch", 5: "movies", 6:"also", 7: "football", 8:"games", 9:"Mary", 10:"too"}

This dictionary has 10 distinct words. And using the indexes of the dictionary, each document is represented by a 10-entry vector as follows:

[1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

In English, there are spaces between each word. However, in Japanese, every words are attached each other. Therefore we need to separate them into pieces of word or other units. We used two major methods to make dictionaries and vectors of documents data.

The methods are as follows:

1. Using words segmented by a morphological analysis
2. Using character *n*-gram

We will explain each approach and their advantage and disadvantage.

4.2.1 Using words segmented by a morphological analysis

The first approach is to use words that are segmented from character strings using a morphological analyzer. We used Mecab[5], a Japanese morphological analyzer.

Advantage

This method can use words in an ordinary sense to make vectors. In addition, it is useful to decrease sizes of indice and dimensions of vectors.

For example, "begin", "began" and "begun" are different forms of the same word, "begin," in English, and using a morphological analyzer, we can easily put these different forms into a single word.

Disadvantage

If some words are not registered in the dictionary of the morphological analyzer employed, these words cannot be analyzed correctly. The information of words are lost.

4.2.2 Using *n*-gram

A character *n*-gram expression is a string of *n* characters. For example, if *n* is set to two, a word "nature" is separated to "na", "at", "tu", "ur", and "re" as 2-gram expressions.

Advantage

We can use this method in any written languages without complex preprocesses.

Disadvantage

An *n*-gram expression may not correpond to a meaningful unit in a usual sense. In addition, the dimension of the vector is very high by this method.

Online customer reviews usually contains expressions that are not registered in a dictionary for morphological analyzers. So if we take the first approach, there is a risk that expressions wrongly segmented badly affects our analysis. On the other hand, if we take the second approach, we need pruning many *n*-gram expressions of lower frequencies to enable a realistic computation. In other words, both approaches have some difficulties. So we compared these two approaces to know which is better for our purpose.

5 EXPERIMENTS AND DISCUSSIONS

5.1 Experiments and results

First, we collected the reviews of the accommodations in Japan from "TripAdvisor". We used only the data which have the tag of "Solo", "Couples", "Family" and "Friends". And we defined two classes that a class with the tag of "Solo" and another with the tags of "Couples", "Family" and "Friends".

We extracted 10,000 data at random from "TripAdvisor". These data are labeled by the two classes. And we divided those data into 5,000 training data and the 5,000 test data.

In the vectorization, vectors are constructed by three data set: morphological analysis, bigram, and trigram that are described in Chapter 4. Moreover we let vectors pass in the filters and made other data sets. The purposes of using filters are to reduce the element of an unnecessary vector and to raise the accuracy of analysis. We thought that heavily used words should have appeared in many reviews. On the other hand, the less frequency words are not special data. Therefore we expected that the accuracy of classification would go up if those words can be removed. We tried to use some simple filters that considered in the

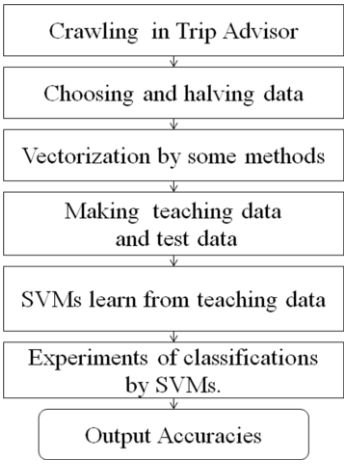

Fig.3. The flowchart of expriments

frequency of words. Seven data sets were made from each method.

- Data containing all the elements
- Data in which deleted the elements that frequency are top 10% , 20% and 30%
- Data in which deleted the elements that frequency are lower 10% , 20% and 30%

We used SVM as the method of machine learning. We used C-SVC of Soft margin classification among the kinds of SVM. Soft margin classification is the technique which allows some errors, loosens restrictions and raises accuracy.

Fig.3 shows the flow of the proposed method.

S1: crawling in Trip Advisor and gathered reviews and their options of tags.

S2: choosing the data of reviews that contain the tags, halving data and making data of teaching data and test data later

S3: making vector from data by the methods.

S4: reduceing elements of vector by the frequency, labeled vector from tag of reviews make teaching data and test data.

S5: SVMs learn from these teaching data.

S6: Experiments of classifications by SVMs using those test data.

Table.1 shows experimental results.

Table 1 Precision Rate of Experimental Results

			word	bigram	trigram
Without pruning			63.8	64.5	65.3
Pruning	top	10%	64.0	64.7	64.5
		20%	64.0	64.1	63.9
		30%	64.2	63.9	64.1
	bottom	10%	56.7	61.6	61.2
		20%	51.9	60.3	60.6
		30%	52.3	59.7	59.7

5.2 Discussions

Most accuracies rates were around 60%. When we use the data in which the top 10%, 20% and 30% of elements are pruned, there is not big difference in precision.

However, when we use the data in which The low 10%, 20% and 30% of elements are pruned, there is a small difference in precision. When we use the data made using word segmented by a morphological analyzer, the accuracy of an experimental result is worse than others. Especially, when the data in which the low 20% are pruned, the result is the worst of all results. It may be because the data whose atomic elements are words segmented by a morphological analyzer contain some important words in their low frequency words. Therefore we must not prune these data and we should give more weights to these data. Moreover, the big difference in accuracy was not observed by *n*-gram.

In this experiment, we cannot confirm the high effectiveness of filtering by the experimental results of the precisions. The precision rates of the data of morphological and bigram increased a little. And, in this experiment, precisions of the data of *n*-gram are not decreased very much. If the data of *n*-gram are used for construction of vectors, filtering technique may be useful for reducing computation times of vectors without decrease of precisions when using the data of *n*-gram.

Considering these experimental results, the experimental results are not of very high quality. We obtained two problems concerning these experiment results. Firstly, we should devise techniques to improve the accuracy rate higher. Next, we devise again techniques to make appropriate filters, especially for the data of morpheme. Filtering is necessary to reduce the computing times of vectors. However, if we apply the filters to the data, the accuracy tends to fall down. Therefore, we must increase the accuracy rate higher before applying the filters.

6 CONCLUSION

We devised a method to classify reviews by using features added by reviewers and we applied the machine learning method SVM to this problem. This work is a preliminary one, and we made a comparison of two data sets whose atomic units are different. Based on this result, we are planning to make an extensive analysis of online customer reviews to construct the best filter set for a certain purpose, and make a better analysis to obtain an intriguing result with online customer reviews.

REFERENCES

[1] Iida R, Kobayashi N, Inui K, Matsumoto Y, Tateishi K and Fukushima T (2005), A Machine Learning-Based Method to Extract Attribute-Value Pairs for Opinion Mining, IPSJ SIG Technical Report, 2005-165-4, pp21-28, 2005.

[2] Hashimoto T, Murakami K, Inui T, Utsumi K and Ishikawa M (2008), Topic Extraction And Social Problem Detection Based On Document Clustring, vol.5, pp216-226, 2008.

[3] Ikeda D, Takamura H and Okumura M (2008), Semi-supervised Learning for Blog Classification, IPSJ SIG Technical Report, 2008-4, pp.59-66, 2008

[4] V. N. Vapnik , The Nature of Statistical Learning Theory, 2nd ed., Springer-Verlag, New York,2000.

[5] http://mecab.sourceforge.net/