Class Project Book Chapter
Innovations in Information System
Series: 3 | Year: 2017 | ISBN: 978-967-0194-97-4

20

EXTRACTING ATTRIBUTES OF A PRODUCT FROM ITS REVIEWS

Seyed Hamid Ghorashi, Roliana Ibrahim, Shirin Noekhah

20.1 INTRODUCTION

Online shopping and selling products on the web has become a popular way of business in recent years. Most of the manufacturers take this chance to advertise their merchandises on a virtual environment where is everyday visited by the people around the world. Thus, they can both engage people to make a review on their products and sell them at the same time. Sometimes, popularity of a product or a particular characteristic of that product leads to receive a large number of reviews from the customers. Many reviews can be too long and only contain a few sentences that deliver the customer opinion. Furthermore, it may be important for the manufacturers to know which attribute(s) of their product has been criticized by the customers. Identifying products attribute is a challenging area of research which has been paid attention by the researchers in recent years.

20.2 OPINION MINING

Opinion mining or sentiment analysis is a technique to track the mood of people about a particular product. It involves building a system to gather and examine opinions about the product made in blog posts, reviews, comments or tweets. Automated opinion mining often uses machine learning techniques that are component of artificial intelligence.

Sentiment analysis is useful in several ways. For instance, in marketing, it can help you to decide for a purchase and determine which versions of a product or service is more popular. Knowing other’s opinion about a particular feature of a product may also be interesting for the purchaser. In particular, a review might roughly be positive about a hand phone, but be specifically against it while we are talking about its weight. Collecting this kind of information in a systematic way, that basically originates from the public opinion, gives a brighter view than surveys which usually do not cover all the aspects.

Opinion mining systems are often implemented by the software that are capable of extracting knowledge from the data stored in a repository. The mining process can be as simple as learning polarity and sentiment of the words, or as complicated as performing deep parsing of data to identify grammar and structure of the sentences.

20.2.1 Extracting Attributes of a Product

The first step in mining customer reviews of a product is to find out whichattributes or features have been paid attention by the customers. After taking the features we can go through with the next step and identify the orientation of the customers. Thus, it will corrupt the process of decision making in the business if the features are identified incorrectly. It is a fact that people express their idea in different manner. The product reviews may contain lots of things that are totally unrelated to the product features. Sometimes also people do not point to the feature directly while they are talking about it. To give a brilliant perspective of what we mentioned above, let us see an example from a review of a cellular phone:

Sentence 1: “The menu options are uncreative.”

Sentence 2: “It’s not easy to carry.”

In the first sentence, the customer is unsatisfied with the menu options of the cellular phone and explicitly points to the feature. The second sentence is talking about the size of the phone while it has not been mentioned directly within the sentence. Therefore, finding and extracting features is a challenging task in text mining which opens a distinct area of research.

20.2.2 Identifying Opinion Orientation on Features

It is likely to be significant for the business holders to know that what their customers think about the products and services. Hence, the second task in the process of mining opinions seeks to classify the opinionated piece of text and identify its sentiment polarity. So far, the researches on sentiment mining have different focuses and objectives. However, they can be fallen into three major categories including sentiment words or phrases identification, sentiment orientation identification, and sentiment sentence or document classification.

Sentiment words/phrases identification has also been called opinion words identification by some papers which is the first phase in sentiment mining projects. These words will be considered as the keys for latter sentiment orientation identification. Identifying sentiment orientation of words or phrases emphasizes on the semantic of the words and directly points to the text writer’s opinion, whereas, sentiment identification of words does not mind the semantic of the word.

Current works to identify sentiment orientation are commonly focused on either verbs or adjectives. Hu and Liu [5] applied POS tagging and some natural language processing techniques and extracted adjectives as opinion words. Then they used WordNet to determine the polarity of extracted adjectives. Finally, the orientation of each sentence is predicted based on the dominant orientation of the opinion words participating in the sentence. A similar work was also done by Nasukawa and Yi (2003) to predict the orientation of the sentences. Besides adjectives they considered verbs as they can effect on the polarity of the sentences. In their method the syntactic dependencies among the phrases are analyzed and phrases with a sentiment term that is modified by a subject or modifies the subject term are extracted. Then they look into the sentiment dictionary to find the polarity of the sentiment term when the term is a verb.

20.2.3 Review Summarization

A summary is a text that is produced from one or many other texts and covers all the significant parts of the original text(s) [4]. Text summarization is the procedures that have to be done by the system to make a summary from the text. The output of a summary system can be either an extract or an abstract. When the output is produced from a significant parts of the original texts, the summary is called an extract summary, whereas, an abstract summary is a summary that can be served as the substitute to the original document [8]. Therefore, review summarization can also been categorized under the group of extract summaries as it only focuses on the parts of the review that are likely to be significant for the readers.

So far many works have been done on summarizing review databases. A related work done by Morinaga et al., 2002, extracts opinions from a collection of product reviews gathered by a general search engine. It then mines the opinions and discovers statistically meaningful information.

Hu and Liu [6] mined the features of some products extracted from a collection of reviews and found the opinions regarding to the features. After that they identified the polarity of the opinions and classified the reviews accordingly.

OPINE is an unsupervised information extraction system proposed by Popescu and Etzioni [2], that extracts features and their associated opinions from online reviews. Using a relaxation labeling technique it determines the semantic orientation of opinion words in the context of the extracted product features and classify the reviews based on the sentiment of the sentences.

Another approach that was proposed in 2008 Kokkoras, Lampridou, Ntonas, &Vlahavas, [7] focuses on additional available parameter called metadata regarding to each review, rather than classifying the reviews based on their polarity. It ranks the sentences of multiple reviews and adjusts their importance based on the features such as familiarity of the user with the domain and the usefulness of each review to the other users.

20.3 CHOOSING A TECHNIQUE TO EXTRACT PRODUCT ATTRIBUTES FROM THE REVIEWS

Finding a proper way to extract knowledge from a set of unstructured documents is a challenging area of research. The extracted knowledge here is the product features, about which reviewers have written their opinion. The field of study is divided into two categories, supervised methods and unsupervised methods. Although, supervised techniques are more precise between the two categories but they have to be trained by the human. The method is effective when the documents are not too away in terms of the subjectivity. This means that if we have two datasets, each of which focuses on a particular topic, the training set for them should be different as well. For instance, the opinion words used to express one’s feeling about a movie is different from the situation he is going to talk about the quality of a product. In a movie dataset some words may carry a negative orientation while the same word in a product review dataset can deliver positive orientation. As an example, in the sentence ‘*The story of the movie was too simple*’, the word ‘simple’ shows that the audience was not satisfied with the story and made an unfavorable judgment. In another sentence regarding to a product review, the customer says ‘*The wide touch screen makes it simple to work with the phone*’, which shows a positive orientation for the word ‘simple’. So accumulating a set of terms orientation as the training data may bring about running into trouble with ambiguous words.

Unsupervised techniques can also be fallen into different categories of approaches such as Machine Learning, Artificial Intelligence, Pattern Mining, etc. These methods can be applied for various datasets, because the training procedure is no longer required. The next following sections focuses on pattern mining techniques as a potential solution to solve the above mentioned issue.

20.3.1 Algorithms for Pattern Mining

Pattern refers to a perceptual structure of the data which helps us to discover the knowledge and convert it to a human-understandable structure. Depending on the problem which is defined, different patterns such as frequent patterns, sequential patterns, periodical patterns, and so on may be declared.

Frequent pattern mining tries to discover frequently occurring patterns and trends automatically, with no intervening of the user. Since the first proposal of this data mining task, a lot of follow-up researches have been carried out that resulted to emerge a variety of similar techniques. The approaches of mining frequent patterns can be put into the three categories of candidate generate-and-test (e.g. Apriori), pattern-growth (e.g. FP-growth, H-Mine), and hybrid methods (e.g. DCI-Closed). The classification of the algorithms is based on the criterions such as the way of traversing the search space, speed of algorithm, memory efficiency and so on [1].

20.3.2 The Motivation of Choosing Frequent Pattern Mining Techniques

Usually, when people discuss and give their opinion on the same thing, their words converge. Moreover, a product feature is a noun or noun phrase which can be appeared in review sentences. Given the fact, it can be inspired that the nouns with high frequency can most likelybe considered as feature words. Frequent pattern mining techniques tend to determine multiple occurrence of the same object. So we can take the advantage of such techniques to search for frequent nouns or noun phrases as the potential feature words.

20.4 COMPARISON OF FREQUENT PATTERN MINING APPROACHES

This section provides a comparative study on pattern mining approaches and three common algorithms which can be applied for discovering frequent features of a product in customer reviews. The following sections discuss about the weakness andstrength of each algorithm for identifying product features.

20.4.1 Candidate Generate and Test Approach

In recent years, many pattern mining algorithms have been proposed by the researchers. All the algorithms can be listed as candidate generate-and-test, pattern-growth and hybrid approaches[1]. The Apriori algorithm is a famous algorithm to mine patterns in the database and find frequent itemsets by following the first approach. Given an input item set of length *n*, Apriori tries to generate a set of items of length *n+1* and then check the candidates if they meet the support threshold [3]. This strategy leads to appear some major limitations. First, by increasing the number of items in the data set, the number of candidate itemsets that should be generated and tested is exponentially increased. Moreover, to test the minimum support of the candidates it is required to scan the whole database over and over again. Consequently, it is a time consuming task which makes the mining process to be slow down. Second, the algorithm relies heavily on using memory resources while generating and maintenance of the candidates. Therefore, it cannot be considered as a proper strategy for the dense data sets.

Apriori algorithm does not take the sequence of the items into account while calculating the candidates support. It means that any combination of two particular items is considered as the same itemset and the support values of them are added up together. This may bring about getting an undesirable result in some cases such as the current study. Let us give an example to clarify the problem. Imagine that $T_1$is a transaction set with four items whose the items are a, f, c, and g respectively. Another two transactions may be defined as$T_2 = \{c, g, b\}$ and$T_3 = \{f, g, c\}$. Given the value of min_support = 2, and running the algorithm, Apriori generates the result as illustrated in table below.

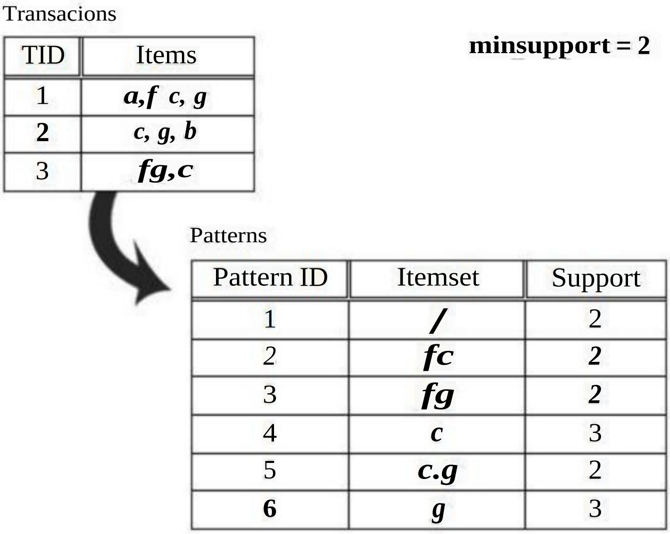
Table 20.1: A sampleresult for Apriori algorithm

Pattern ID	Itemset	Support
1	*f*	2
2	*c*	3
3	*g*	3
4	*f,c*	2
5	*f,g*	2
6	*c,g*	3
7	*f,c,g*	2

From the table, we can see that Apriori does not recognize a distinction between the itemsets {*c*, *g*} and {*g*, *c*} and it counts every possible sequence of two items *c* and *g* to calculate the support. Assuredly, in the domain of market basket analysis it does not effect on the result but when we are talking about mining texts, the order of the words should be taken into account. For example, if we replace the items *c* and *g* with the words ‘video’ and ‘camera’ in a document, knowing the frequently occurrence of both ‘video camera’ and ‘camera video’ noun phrases is important to make decision.

20.4.2 Pattern Growth Approach

As it was mentioned in the last section, the second group of pattern mining approaches is called pattern-growth technique. In this strategy, it has been attempted to remove the above stated limitations of the first approach (candidate test-and-generate) and make the process of mining faster. One of the algorithms called H-Mine (Hyper-Structure Mining of Frequent Patterns) achieved a good result on preserving memory space, and also accelerating the mining procedure. According to the frequent patterns found, H-Mine recursively partitions the database into sub-databases and looks for the local frequent patterns in the new search space [9]. Another privilege of H-Mine is the ability of handling the last issue discussed in the previous paragraph.



Transacions

TID	Items
1	*a,f c, g*
2	*c, g, b*
3	*fg,c*

minsupport = 2

Patterns

Pattern ID	Itemset	Support
1	*/*	2
2	***fc***	***2***
3	***fg***	***2***
4	***c***	3
5	***c.g***	2
6	***g***	3

Figure 20.1 A sample result for H-Mine algorithm

Figure 20.1 shows the result of H-Mine after running it on the same transaction set of previous example. The difference between the results generated by the two algorithms can easily be perceived at the first glance. By looking at the figure we can see that H-Mine differentiates between the two itemsets {*c, g*} and {*g, c*}, and it considers each occurrence of them individually. So the support values are presented as 2 and 1 respectively. Whereas, the minimum support value was already defined as 2, those itemsets with lower support were ignored and only {*c, g*} was presented. This characteristic of H-Mine makes it possible to be used by this study for finding the strings of words that may be occurred frequently in our document repository.

20.4.3 Hybrid Approach

Algorithms adopting this category of mining techniques, inherit the properties of both previously discussed methodologies. An example is DCI-Closed proposed to mine frequent closed patterns in databases. It traverses the search space in depth-first search manner like algorithms of ‘divide-and-conquer’ strategy. However, a single candidate is generated each time like the algorithms exploiting ‘test-and-generate’ technique [1]. Unlike Apriori which suffers from taking many scans of the database while generating candidates, DCI-Closed creates a vertical bitmap representation of the data with only two successive scans. This strategy increases the performance of the algorithm than the Apriori while dealing with large datasets.

DCI-Closed tries to discover all the candidate itemsets that are closed and frequent, while Apriori only focuses on frequent ones. In the domain of text mining there might be some situations in which it is required to search and discover frequently occurred phrases. Every phrase is composed of a sequence of words that can be considered as individual items in a transaction (sentence). In case where the frequency of an itemset (here means a phrase) is exactly the same with its subsets (individual words), it can be realized that the items are most likely to be used together to deliver a particular meaning. Therefore, the subset items are not supposed to be considered as frequent items and only their superset should be identified. Let us give a real world example for better understanding of the issue.

Table20.2 Sample comments from a review dataset

ID	Comments
1	I was looking for a budget **video camera** with a good quality. This can be a good choice and I suggest it to you guys.
2	In low light the quality of *this* ***videocamera*** is poor.
3	The **battery** life for me seems to be awesome. I haven't run into any problems with the **battery**.
4	I was really impressed by the quality of this digital **video camera**.
5	Thanks to its 8GB built-in *flash memory* and 500mA Li-Ion **battery**. *The**camera***can capture *images* continuously for up to 2 and half hours.

Table 20.2 provides some comments from a partial review database. It has been supposed here that every noun or noun phrase with occurrence of at least two times in the whole dataset is considered as a frequent item. By looking at the reviews, it is anticipated that the highlighted words are identified as the frequent feature words. Running each of the three algorithms discussed earlier on the dataset give the following results:

Table 20.3 Extracted features by Apriori, H-Mine and DCI-Closed

	Apriori	**H-Mine**	**DCI-Closed**
Generated Itemsets	video	video	----
	camera	camera	camera

	video camera	video camera	video camera
	battery	battery	battery

It can be found from the table that the first two algorithms are not able to handle the situation above and they return the same results. Therefore, DCI-Closed was applied in the project and precision of the system was measured. The experimental result is discussed in the next section.

References

[1] Ben Yahia, S., Hamrouni, T., & Mephu Nguifo , E. (2006). Frequent closed itemset based algorithms : A thorough structural and analytical survey. *ACM SIGKDD Explorations Newsletter*, 93-104. New York, USA: ACM.

[2] Carenini, G., T. Ng, R., & Zwart, E. (2005). Extracting Knowledge from Evaluative Text., (pp. 11-18).

[3] Hemalatha, R., Krishnan, A., & Hemamathi, R. (2005). Mining Frequent Item Sets More Efficiently Using ITL Mining. *3rd International CALIBER*. Ahmedabad: INFLIBNET Centre.

[4] Hovy, E. (2005). Text Summarization. Oxford University Press.

[5] Hu, M., & Liu, B. (2004a). Mining and Summarizing Customer Reviews. Proceedings of the tenth ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA, 22 - 25 August.

[6] Hu, M., & Liu, B. (2004b). Mining Opinion Features in Customer Reviews. Proceedings of the 19th national conference on Artifical intelligence. San Jose, California, 25 - 29 July

[7] Kokkoras, F., Lampridou, E., Ntonas, K., & Vlahavas, I. (2008). "Sumarization of Multiple, Metadata Rich, Prduct Reviews." Aristotle University of Thessaloniki, Greece.

[8] Lloret, E. (2006). "Text Summarization:an Overview." Spanish Government Report.

[9] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., & Yang, D. (2007). "H-Mine: Fast and space-pre serving frequent pattern mining in large databases." *IIE TRANSACTIONS*, 39(6): 593-605.