Searching Blog Sites with Product Reviews

Hironori Kuwata, Makoto Oka, and Hirohiko Mori

Tokyo City University, 1-28-1, Tamadutumi, Setagaya, Tokyo, Japan
{g1181811,moka,hmori}@tcu.ac.jp

Abstract. Recently, buzz marketing sites gives the information that is useful for consumers and companies. They want to customer feedbacks of feeling and experience. However, the searched results contain huge numbers of commercial sites when user search review with traditional search engine. We search blog site that include review sentence. We need to decision whether document of blog site include review sentence. Thus we think that two process to decide blog site whether review blog site. The first process creates to data set for certain product that viewpoint feature word. In this paper, feature word is tow term in the evaluated perspective word and evaluated value word. Data set is information for making decision sentence whether review sentence. Second process is a search for review sentence. This process decided blog site whether review blog site. This process use extracted opinion tuples from one sentence of blog site document and created data set to decide sentence whether sentence is review sentence. This process decided review blog whether document of blog site include one and more review sentence. We proposed review blog site searching system that system have two process.

Keywords: natural language, opinion extraction.

1 Introduction

Recently, buzz marketing sites gives the information that is useful for consumers and companies. Consumers watch buzz marketing sites such as Amazon.com when they want to buy some items. For consumers the reviews can be used as the reference information in purchasing a product, and for companies they are useful to develop the products of the next generation by watching reputations of their and another companies' products. However, word-of-mouth is difficult to search with traditional search engine. They are looking reviews for some media.

Buzz marketing comments are contained in many review sites, blogs and twitters so on. On the review sites, we can collect a lot of reviews but, the reviews are often short and do not contain much detail information. On the other hand, review blog sites sometimes include more detail information, such as long-term reviews, but traditional search engines by the keywords matching find too much sites. For example, when we want to find blog sites where the reviews about a compact digital camera, the searched results contains huge numbers of commercial sites and we must spend much time to find review blog site among them. Commercial site are online shop site, new

S. Yamamoto (Ed.): HIMI/HCII 2013, Part III, LNCS 8018, pp. 495–500, 2013.

products information and so on. We think that review blog site is an efficient media to get review if we can find only review blog sites.

We propose searching system that detects review blog site. User wanted information for blog site that information is customer feedbacks sentence of feeling and experience. Commercial sites not contain customer feedbacks of feeling and experience. We define a review blog site that contains one or more customer feedbacks sentence of feeling and experience.

2 Related Works

Kuwata[1] proposed method to organize free written review of review site with the use of other review site. How to organize this study is classification review sentence into viewpoint given review sentence of item. Viewpoint is that of camera, such as design, function and battery so on. Typically, training data are created manually when classified sentence. His system substitute reviews of like Viewpoints.com for training data. He created training data automatically extracted review for each viewpoint, as shown in figure 1. These reviews will be referred to as review corpus in this paper. They created data set that is pair viewpoint and feature word. They have extracted feature word from the review corpus that feature word is occurring frequently word only certain viewpoint. He classified review sentence into each viewpoint with the use of data set. He showed review corpus is substitute for the training data.

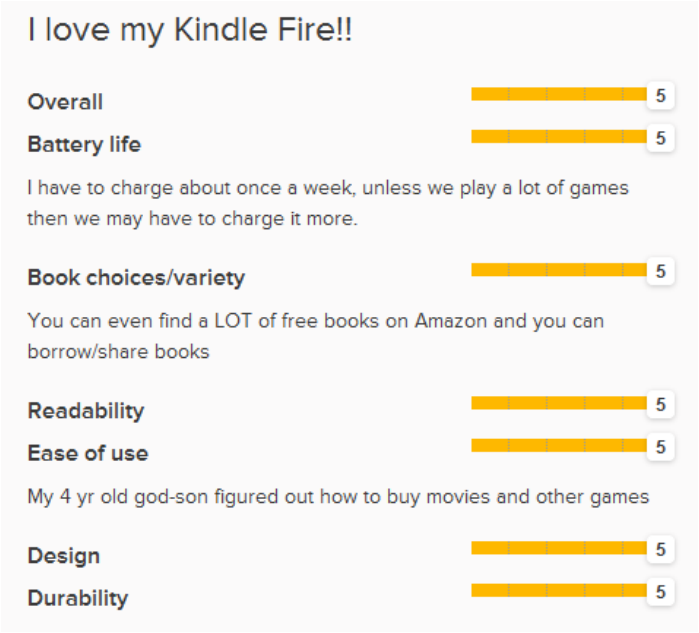

Fig. 1. Viewpoints.com's review

Table 1. Five syntax opinion pattern

1	X→Y
2	(X→Y1)→Y2
3	X1→(X2→Y1)
4	((X1→Y1)→X2)→Y2
5	YX (modification relation)

X:evaluated perspective word Y:evaluated value word →: dependency relation

Sugiki[2] proposed new presentment of search results of accommodation review site that method to sort from does not quantification point of view. His system sorts review in descending order of relation between user query of natural language and review sentence. His measured the relevance of opinion tuples (evaluated perspective word and evaluated value word). His system extracted opinion tuples with the use of syntax opinion pattern. Using syntax opinion pattern can be extracted opinion tuples when review sentence had the same sentence structure as five syntax opinion pattern (Table 1.). His study showed that these five patterns can use extracted opinion tuples. However, using syntax opinion pattern has a problem. Syntax opinion pattern considers the category of sentence in order to extract the structure of sentence. His study has solved this problem through abounded scope for search page.

In this paper searched review blog site from searched results by traditional blog search engine. We think that it is difficult to use a syntax opinion pattern for blog site. We adapt to blog site for combination of syntax opinion pattern and data set.

3 Searching System

In this paper, we proposed system that searched review blog site that is contained one or more review sentence in blog site. Our system classified blog site that are extracted by traditional blog search engines whether review blog site. We create a system that determined one sentence whether review sentence (Fig.2). Processing of this system consist of two processes. The first process creates to data set for certain product that viewpoint and opinion tuples from review corpus. This process is a preparation to classify one sentence. Opinion tuples are two term in the evaluated perspective word and evaluated value word. Data set is information for making a decision sentence whether review sentence of blog site. Thus we choose opinion tuples for product feature. Second process is a search for review sentence. This process decided whether sentence of blog site is review. This process use extracted opinion tuples from one sentence of blog site document and created data set. Discrimination circuit of this process classified sentence whether review sentence.

We determine blog site when second process classified one or more review sentence from blog site documents. This study targets domain of the digital single-lens reflex camera. Sentences are pre-multiplied by morphological analysis for the words to deal with Japanese.

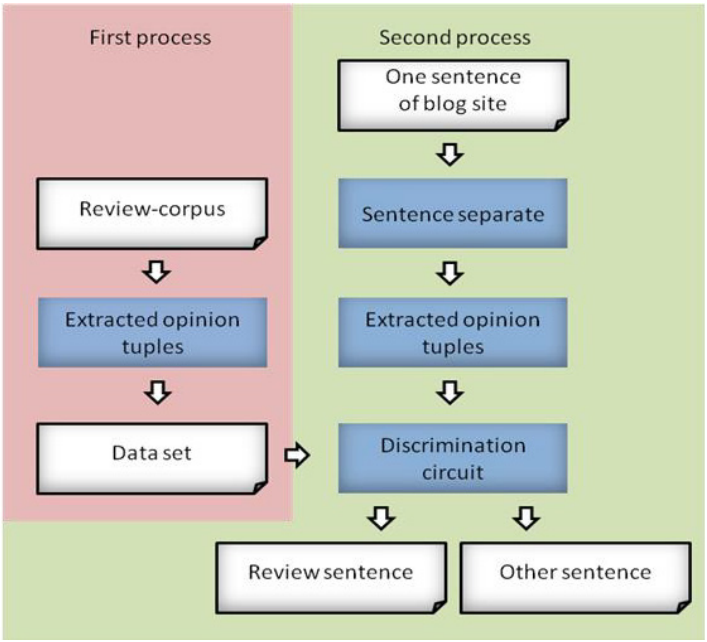

Fig. 2. System to decide whether review sentence flow

3.1 Data Set

We create data set of the product in order to determine whether review sentence. This study can't be to extract data set of the production from the general corpus. We need corpus that written about a certain product. However, handmade corpus is costly alternative. Review corpus is low cost because it can be generated automatically that review corpus written for each viewpoint. We create data set from review corpus.

Our system extracted data set from one sentence of review corpus with the use of five syntax patterns. Data set have viewpoint and opinion tuples (evaluated perspective word and evaluated value word). Review corpus is review sentences of review site written for each viewpoint. Our system extracts the opinion tuples one sentence for each viewpoint. Opinion tuples are extracted when one sentence has the same syntax as the five patterns. Our system checks sentence for sentence structure with the use of syntax analysis. Our system extracted opinion tuples when the sentence has same sentence structure as one of five syntax opinion patterns. Our system erases unnecessary word class from extracted opinion tuples. In this paper, view point is "Design" , "Image quality", "Ease of use", "Battery", "Portability"," Function", "LCD" and "Grip" of digital single-lens reflex camera review corpus. These viewpoints have opinion tuples that have been extracted in the above.

We define that feature word of certain product is like "functional" and "button" (Table 2.) We calculate the frequency of these extracted viewpoint, evaluated perspective word and evaluated value word. Word set is a collection of high frequency only in a viewpoint.

Table 2. Frequently-appearing evaluated value word

		Viewpoints								Total
		Design	Image quality	Ease of use	Battery	Portability	Function	LCD	Grid	
evaluated value word	think	1089	1516	1166	851	925	1175	705	694	8121
	big	136	302	438	529	1283	1392	1719	2301	8100
	good	705	981	658	489	497	605	402	620	4957
	shoot	45	932	475	1111	138	1214	442	212	4569
	use	125	578	620	632	166	725	257	136	3239
	functional	120	104	533	52	120	1823	98	42	2892
	feel	321	619	352	193	363	313	196	326	2683

3.2 Classification Review Blog Site

Our system classified blog site that extracted by traditional blog search engine whether review blog site. We define a review blog site that contains one or more review sentence. We create a system that determined one sentence whether review sentence.

Our system need to perform a process separated one sentence document of blog site. There is a one sentence that did not end with a period Japanese document of blog site. For example, one sentence ends with line breaks. Japanese used verbs such as “desu” and “masu” so on at the end of case one sentence. We consider sentence end whether sentence ends verbs and line breaks.

Our system extracted opinion tuples (evaluated perspective word and evaluated value word) with the use of five syntax opinion pattern from separated one sentence. Opinion tuples are extracted when one sentence has the same syntax as one of the five patterns. Our system checks sentence for sentence structure with the use of syntax analysis. Our system extracted opinion tuples when the sentence has same sentence structure as one of five syntax opinion patterns.

The sentence comes up for review sentence that had same sentence structure as the one of five syntax opinion pattern. Sentence of blog site mixed certain review sentence and other sentence.

We define that review sentence when it is discovered that there is opinion tuples of the digital single-lens reflex camera. This system decides opinion tuples whether about the digital single-lens reflex camera with the use of data set. This system determine review sentence when the same extracted opinion tuples and data set. Our system determined to be a review blog when document of blog site include review sentence.

4 Discussion

Sugiki[2] use syntax opinion pattern on sentence of review site. We have not been able to confirm that syntax opinion pattern adapt to free description sentence. We verified effect of syntax opinion pattern to extract opinion tuples of free description sentence. This study use sentence that mixed review sentence and other sentence. We need to determine five syntax opinion pattern can be extract evaluated perspective and evaluated value of the digital single-lens reflex camera.

Test data is free description review on Amazon.com. This data include 250 review sentence and 250 other sentence. We have verified manually to decision extracted opinion tuples whether evaluated perspective and evaluated value word. We get the results of precision rate 93.5% and recall rate 62.8%. Although it is at the discretion of the manually, we think that our system can be classified review sentence with data set and extracted opinion tuples of blog site. Recall rate is low because the sentence is not the correct structure. We think syntax opinion pattern work poorly when reviewer written sentence of very short and spoken language. Many short sentence is itemize as ”DESIGN: VERY NICE!!“. Many of the spoken language are not subject word and mistake grammar. Japanese spoken language is often omitted the subject word. We need to think of solution for sentence is not subject word.

As it stands, data set is viewpoint, evaluated perspective word and evaluated value word. We expect the results of high compliance rate and low recall rate with the use of its data set. We should consider new data set to determine many review sentence.

5 Conclusion

In this paper, we have proposed system that searched review blog site. Our system classified blog site that are extracted by traditional blog search engines. We create a system that determined one sentence whether review sentence. Sentence of blog site contain many sentence other than review unlike document of review site. Thus we solve two processes. The first process creates to data set for certain product that viewpoint and opinion tuples from review corpus. Data set is information for making a decision whether sentence of blog site is review. Second process is a search for review sentence. This process use extracted opinion tuples from one sentence of blog site document and created data set to decide sentence whether sentence is review sentence. We think system can search for review blog site if there are two processes.

6 Feature Works

We have three challenges. First, we need examine again the effect of such as very short sentence and spoken language because we deal with free description in this study. Second, we create new data set. In the current word set is expected to increasingly not given review sentence. Third, we need to create a new classifier when the foregoing has been improved. We think to be classified in two stages. First stage, system classifies certainly review sentence. Second stage, system classifies dropped review sentence of first stage with the use of new data set.

References

1. Hironori, K.: An evaluation viewpoint distinction automatic classification of a review. IPSJ Processing, 2.399–2.400 (2011)
2. Sugiki, K, Matsubara, S.: Natural language search based on information extraction from the reputation of review sentences. IPSJ SIG Technical Reports, Research Report of Natural Language Processing (124) (2006)