

Mining Relationship Between User Purposes and Product Features Towards Purpose-Oriented Recommendation

Sopheaktra Yong^(✉) and Yasuhito Asano

Department of Social Informatics, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
sopheaktra.yong@outlook.com, asano@i.kyoto-u.ac.jp

Abstract. To help in decision making, buyers in online shopping tend to go through each product features, functionalities, etc. provided by vendors and reviews made by other users, which is not an effective way when confronting loaded of information and especially if the buyers are beginner users who have limited experience and knowledge. To deal with these problems, we propose a framework of purpose-oriented recommendation which present a ranking of products suitable for a designated user purpose by identifying important product features to fulfill the purpose from user reviews. As technical foundation for realizing the framework, we propose several methods to mine relation between user purposes and product features from the online reviews. The experimental results employing reviews of digital cameras in Amazon.com show the effectiveness and stability of proposed methods with acceptable rate of precision and recall.

Keywords: Recommendation system · Review analysis · Bootstrapping

1 Introduction

Nowadays, increasing large numbers of consumers have gradually shifted their purchasing behavior to online shopping sites. Online retailers like Amazon has set up a platform where vendors could upload their products together with descriptions illustrated the functions and features of the product. Consumers often browse through other users' feedbacks to help in their decision-making. However, it is tough sometimes to go through all the reviews since some products might have hundred comments. Many sites have developed various recommender systems to help users choosing suitable products for their needs. The common techniques are based on the similar products a consumer used to buy previously or based on other consumers who share similar interest or purchasing history. However, it is still not sufficient for users since the system does not recommend them according to their needs particularly when it comes to the commodities designed for variety of usage purposes such as PC, smartphone, digital camera etc. For instance, there are various purposes associating with digital camera such as shooting portrait, landscape, wildlife, sports, car racing, etc. Generally, different product features are required for different purposes. Camera makers could not describe

them all in product description or advertisement. Various recommended techniques with an emphasis on product features including methods to automatically extract product features from the reviews and techniques for ranking products based on features have been proposed in [1–5]. However, it is still very difficult for novice user to choose the suitable camera since they could not tell which camera features are good for their specific needs.

For this reason, we propose a framework of purpose-oriented recommendation which can greatly improve the convenience by recommending products automatically based on consumer’s purposes. In this paper, we describe a methodology to mine hidden knowledge from the online reviews, particularly, to extract relationship between user purposes and product features as the technical foundation to realize the proposed framework.

Our Contribution. The contribution of our paper can be summarized as follows:

- We propose a purpose-oriented recommendation framework to suggest products suitable for users especially novice users who do not have much knowledge about the products and a methodology as the technical foundation for realizing the purpose-oriented recommendation framework.
- We present a bootstrapping-based methodology employing state-of-the-art Natural Language Processing (NLP) techniques to extract the relationship between user purposes and product features from online reviews.
- We Introduce (1) a conditioning approach for filtering new seeds by computing the word semantic similarity to lessen the common problem of bootstrapping – semantic drift and (2) an ontology-based technique to generate more seeds for bootstrapping process.
- We tested and confirmed the effectiveness of the proposed methods through experiments on very popular and large Amazon review dataset and user evaluation.

The rest of the paper is organized as follows: our framework of purpose-oriented recommendation is illustrated in Sect. 2. Section 3 presents the methodology for extracting the relationship between product features and user purposes. Section 4 contains experimental results. Brief reviews of related works conducted in this area are summarized in Sect. 5 followed by the conclusion in Sect. 6.

2 Purpose-Oriented Recommendation Framework

As we described above, the users tend to face difficulties when deciding which products they should purchase for their specific needs if the products contains lots of technical features and functionalities. Figure 1 illustrates the flow of the recommendation framework we wish to build in future work.

First, the system will present a list of all possible purposes for a particular product type, e.g. digital camera. In this study, we assume that the users are aware of their purposes and asked to choose from one or more purposes from the purpose list presented by the system. After that, the system will output the features which are most relevant and associated with the chosen purposes. Here, the system also analyses the optimal or minimum requirement for each product features to achieve the

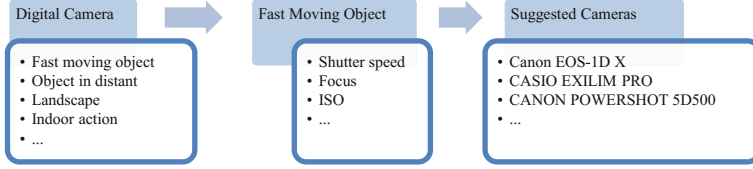


Fig. 1. Purpose-oriented recommendation framework

corresponding purposes. For example, at least a burst rate of 6 frame-per-second is required to capture a flying bird. These kinds of information could help users to not only pick the right camera but also educate them to set the right setting for shooting a specific target. Finally, according to the relationship between product features and purposes, the system will recommend suitable products to the users.

In order to realize the framework, the most crucial part of the system is to find the association between product features and purposes which expressed in the users' reviews. Next section, we will describe the methodology in detail.

3 Methodology

The proposed methodology is composed of three main components, Component I - NLP Preprocess, Component II - Initial Seeds Generation, and Component III - Purpose-Feature Relationship Extraction Bootstrapping Model. The method extracts keywords representing purposes from the online reviews and connect them with product features. In this study, product features as the input are manually selected. They can also be extracted from several online shopping websites such as cnet.com, dpreview.com, amazon.com, etc. There have been several studies on identifying and extracting product features automatically, which is not the focus of this research. Another input of the system is text data of consumer reviews. In Component I, the customer reviews are preprocessed using sets of NLP tools. After that, purposes are extracted using a combined technique of Labeled LDA [6] and Word2Vec [7] by Component II. Then starting with a few initial seeds, Component III iteratively learns patterns to extract more seeds. Figure 2 illustrates the overview of our method.

3.1 NLP Preprocess (Component I)

NLP is applied for linguistic preprocessing and raw text analysis such as removing stop words, sentence tokenizer, and part-of-speech tagging. The tool used in this paper is NLTK: The Natural Language Toolkit [8]. A typical review contains several sentences. Here, we employ Punkt Sentence Tokenizer for splitting reviews into sentences.

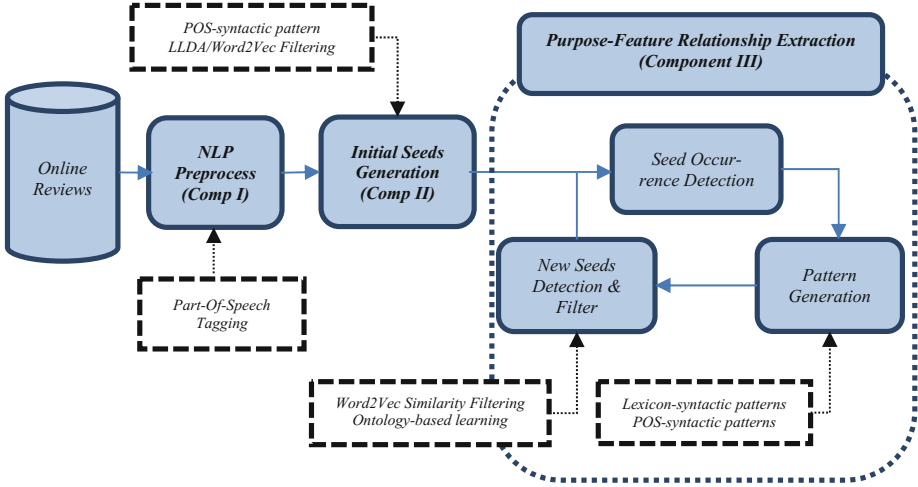


Fig. 2. Overview of methodology

3.2 Initial Seeds Generation (Component II)

Choosing of seeds is arguably the most critical step in bootstrapping. However, most of the previous researchers either chose manually [9] or picked the most frequently occurring words in their corpus [10] that they have identified belong to the category. In this study, we propose a technique which combine topic modeling and word embedding technique to generate initial seeds for bootstrapping model. Our initial seeds generation process is composed of three steps: candidate sentences extraction, term extraction, and term filtering.

We assume that consumers described their usage experiences or purposes inside the reviews. They often exemplified how they use the product or specific function to perform or to achieve something. Table 1 shows some sample reviews containing intended use or purposes of digital cameras' shutter speed functionalities. Mining this type of reviews would enable us to detect the purposes expressed by the consumers. We find "candidate sentences" which contains product features and purpose-oriented expression, by the procedures explained below.

The process of generating initial seeds is described as follows:

1. **Candidate Sentences Extraction:** Candidate sentences are sentences containing product features as well as purpose-related expression. Sentences containing about product features often have expression whether such features are good or bad for a particular purpose. For example, there is a statement - *"The unlimited continuous shooting mode, which is perfect for catching fast-action sports shots"*. Here, the phrase *"Catching fast-action sports shots"* is the purpose of using the *"shutter"* feature. This piece of review tells us that if we consider buying a camera for shooting a fast-moving object, we should choose the camera with the high *"shutter"* speed. To detect the candidate sentences from the reviews, we manually define some syntactic-sentence patterns used to express purposes. An example of the pattern is

Table 1. Customer reviews from [Amazon.com](https://www.amazon.com). Product features and corresponding purposes are highlighted.

Review #1: The unlimited continuous shooting mode, which is perfect for catching fast-action sports shots, solves the problem of slow shutter speeds and is not found in any of the other camera's in the s3's class. Feature: shutter, Purpose: catching fast-action
Review #2: I especially liked the panoramic feature for the beautiful Alaskan landscapes and the burst feature for shooting wildlife in action. Feature: shutter, Purpose: shooting wildlife
Review #3: The shutter speed allows her to take pictures of friends in motion [dancing, skiing and other sporting events] amazing window on back for best picture view. Feature: shutter, Purpose: dancing, skiing, sporting events

<feature> + Verb + Adjective + Preposition + Noun/Noun Phrase. Then we build the set of candidate sentences by extracting any sentences following these patterns.

2. Term Extraction: Once we have a list of candidate sentences, we extract the “purposes” from this set using topic modeling – Labeled LDA. Unlike the conventional LDA, Labeled LDA is a generative model which constrains the topic model to use only those topics corresponding to a document’s label set [6]. In our setting, a list of product features with their corresponding candidate sentences are prepared to train the model. The product features are trained as the topics (labels) and the candidate sentences are treated as documents. Thus, The Labeled LDA outputs the important terms per product feature. These terms are considered as keywords or key phrases representing purposes. An example of this component output is {shutter speed: kids, wildlife, action_shots, recording, getting}, where “shutter speed” is the product feature (topic) and “kids, wildlife, action_shots, recording, getting” are the top keywords representing purposes.
3. Term Filtering: The list of extracted keywords might contain unrelated words since the Labeled LDA output the list based on word frequency only. Therefore, we propose a filtering approach by calculating word semantic relationship using a recent technique Word2Vec. It is a two-layer neural net that processes text. It can convert words into vectors by learning a corpus. Word2Vec can compute the word semantic similarity by calculating their vectors. The corpus we used to train Word2Vec is the consumer reviews. In our methodology, we utilize Word2Vec to filter out the anomaly words from the Labeled LDA’s output. For example, Word2vec is able to remove the unrelated words - “recording” and “getting” from the Labeled LDA output above.

3.3 Extracting Purpose-Feature Relationship (Component III)

To obtain more purpose-related sentences or purposes, we adopt a bootstrapping method as in [9] by detecting more sentence patterns for describing purposes to extract appropriate nouns and noun phrases. Because a simple bootstrapping method tends to produce noises in the results, we also proposed filtering methods to reduce those noises.

Our bootstrapping mainly iterates between the following two phases: (1) pattern generation and (2) seed extraction and filter. The algorithm begins with initial seeds and then iterates through the phases until it could not extract any more patterns or generate new seeds with the word similarity compared to the seeds pool higher than a given threshold τ_2 .

Pattern Generation. In the pattern generation phase, we would like to detect the extraction patterns in the form of lexicon and POS pattern which connect between product feature and user purpose in the reviews. The process starts with a few feature-purpose seed tuples obtained from Component II. For every seed tuple $\langle x, y \rangle$, e.g. $\langle x = \text{shutter}, y = \{\text{kids}, \text{wildlife}, \dots\} \rangle$, we first retrieve all the sentences containing these terms x, y . Next, all “connecting text” linking terms x and y (*kids| wildlife|...*) are extracted into the candidate list. Since we also consider detecting POS syntactic patterns, the “connecting text” is parsed and tagged into the part-of-speech. As shown in the first sentence in Table 2, the phrase “*is good for*” (or POS-tagged “*VB JJ IN*”) is the pattern connecting between product feature (*shutter speed*) and purposes (*kids*). The strength of each pattern is computed by how frequently they are used to connect the seed tuples. Finally, we augment the extraction pattern if any pattern in the candidate list appeared more than a specific threshold τ_1 in the reviews.

Table 2. Sample sentences and extracted patterns

Sentences	Lexicon patterns	POS patterns
Shutter speed is good for kids	<i>is good for</i>	<i>VB JJ IN</i>
I use burst feature for shooting wildlife	<i>for shooting</i>	<i>IN VBG</i>

Seed Extraction and Filter. Most common bootstrapping models are pattern-based approaches in which new seeds are accepted based on the extraction patterns. This could lead to a well-known flaw in bootstrapping known as “semantic drift”. To avoid this problem, we adopt Word2Vec word similarity as the filtering methods so that the current seeds do not wander away from the original semantic meaning of the initial seeds. The process is simply described as following;

- First, our method retrieves from the corpus the set of words W that match any of the extraction patterns P , e.g. Shutter speed is good for “**sport events**”. The word W could be expressed in multi-word terms as Noun or Noun Phrase (NP).

$$NP : \{ (\langle VBG \rangle | \langle JJ \rangle | \langle JJR \rangle | \langle JJS \rangle) * (\langle NN \rangle | \langle NNS \rangle) + \}.$$

- The term W is augmented to the seed list only if its similarity exceeds the predefined similarity threshold τ_2 in average.

In the next section, we will introduce an ontology-based technique to generate more seeds in complement to the conventional pattern-based bootstrapping approach.

Ontology-based Approach for Seed Extraction. As described in previous section, we could extract new seeds or purposes only if the terms are expressed in the extraction patterns. Therefore, we could miss some interesting purposes because not all the customers would use the same sentences to express their purposes or experiences. Let's look at the 3rd review in Table 1. The terms “*dancing, skiing and other sporting events*” share the hyponym relation which means that if one of them is classified as the purpose, so does the rest. For instance, the context pattern of “*allows her to take pictures of friends in*” is not an extraction pattern, thus the terms “*dancing, skiing and other sporting events*” would not have a chance to be detected. In this study, we propose an additional approach and in complement to the pattern-based bootstrapping in detecting new seeds or purposes by mining the new purposes if they have taxonomic relations with any in the seed list.

Taxonomic relations are the most important semantic relations in a domain ontology, the extraction of which has been well studied in the field of lexicon building. We adopt the lexicon-syntactic patterns [11] for taxonomic relation extraction. A total of five lexicon-syntactic patterns are used and described in Table 3. The process of extracting taxonomic relations and generating new seeds for bootstrapping is described as following:

Table 3. The lexicon-syntactic patterns of taxonomic relation extraction

1. NP_0 such as $NP_1 \{, NP_2, \dots \text{ (and/or) } NP_n\}$	Hyponym (NP_i, NP_0)
2. NP_1 is a kind of NP_0	Hyponym (NP_i, NP_0)
3. $NP\{, NP\}^*\{, \}$ or other NP_0	Hyponym (NP, NP_0)
4. $NP\{, NP\}^*\{, \}$ and other NP_0	Hyponym (NP, NP_0)
5. NP_0 , including $\{NP\}^*$ or/and NP	Hyponym (NP, NP_0)

- First, we (1) retrieves from the corpus the set of NPs that match any of the pre-defined patterns and (2) scan whether at least one of the NP set is already in the seed list.
- Next, the rest of the NP in the set is considered as seed candidates and are augmented as new seeds if its word semantic similarity exceeds the predefined similarity threshold τ_2 in average against the other terms in seed list.

Adopting this approach could help us extracting both more precise and general purposes which is more convenient for novice users to comprehend the meaning. It could also assist us to classify the purposes into hierarchy straightforwardly.

4 Experimental Results

In this section, we will illustrate our experimental results and case studies. We conducted our experiments on customer reviews on digital cameras from Amazon.com. The dataset is made available via SNAP, Stanford University [12]. The dataset contains 203,773 reviews and were collected from May 1996 to July 2014.

4.1 Extraction Purpose-Feature Relationship Using Bootstrapping Approach

In this experiment, we manually select 8 camera features and analyze the performance of each model separately by evaluating the number of purposes it could generate, precision, and (relative) recall of all various models i.e. Lexicon, POS, LP, O, and LPO. The first three models are the conventional pattern-based whereas the LPO is the proposed model which combine the pattern-based and ontology-based together. Each model description is as follow:

- Lexicon: Generic patterns are only learnt from lexicon-syntactic patterns
 POS: Generic patterns are only learnt from POS-syntactic patterns
 LP: Generic patterns are learnt from new seeds generated from lexicon and POS syntactic patterns per iteration
 O: Generic patterns are ignored in this context. New seeds are only generated from the ontology-based approach (the hyponym relation in Table 3)
 LPO: A combination of all models

Table 4 shows initial seeds for each camera feature and its selective corresponding new seeds generated by the proposed model – LPO. Note that some camera features’ initial seeds are manually modified prior the bootstrapping process. Table 5 shows the total number of new seeds, precision, and relative recall generated by each model separately. LPO could generate the most seeds. Particularly, a combination of Ontology learning and LP increase the number of seeds than individual model. It concludes that by discovering new seeds via ontology relation during the bootstrapping process could increase the number of seeds accordingly.

Table 4. Initial seeds used for each camera feature and outputs from proposed model. The incorrect purposes are bold.

	Shutter	ISO	Weight	IS	Screen	Flash	Pixel	Focus
Seeds	kids sports wildlife	night light bright	vacation trip carrying	wildlife capturing	reviewing previewing low_angle	light indoor night	printing uploading	kids macros moving
Output	skiing action shots baseball sporting pets...	light condition dim lit darkness dusk...	bicycle_ touring hiking adventures walks...	kids animals fish candids sports...	angle viewing inwards eye_piece composing framing...	castle dim_light dark_room museums ...	publishing share transfer editing email...	moving_targets insects portraits action shots...

Although it is difficult to get all the correct instances (all purposes) for each camera feature, it is still possible to compare the recall of a method relative to another method recall. Following the Pantel et al. [13], the relative recall $R_{A/B}$ of method A given method B can be calculated as:

Table 5. Precision, relative recall, and F-score of each method

Methods	#New seeds	Precision	Rel recall	F-score
Lexicon	94	0.84	0.40	0.54
POS	240	0.72	0.87	0.79
LP	235	0.71	0.84	0.77
O	12	0.80	0.05	0.09
LPO	263	0.75	0.99	0.85

$$R_{A/B} = \frac{R_A}{R_B} = \frac{\frac{C_A}{C}}{\frac{C_B}{C}} = \frac{C_A}{C_B} = \frac{P_A \times |A|}{P_B \times |B|}, \quad (1)$$

where R_A is the recall of A , C_A is the number of correct seeds extracted by A , C is the (unknown) total number of correct seeds in the corpus, P_A is A 's precision in our experiments, and $|A|$ is the total number of seeds discovered by A . We can compute the relative recall by comparing the number of seeds extracted by each model with the number of seeds retrieved by all the models together. The total number of seeds is 272 keywords in which 199 keywords are corrected. We asked a professional photographer who has experiences for more than 10 years to annotate and evaluate the correctness of our extracted purposes. Therefore, we could calculate the performance for all models as shown in Table 5. The lexicon model generates a relatively small number of purposes (94 keywords) with the best precision while the POS and LP model could generate much more purposes (235–240 keywords) but the precision is dropping. It might be implied that we could get more purposes but the it might also contain a lot of noises in the output. However, our proposed model (LPO) which combines pattern-based and ontology-based together can not only generate more purposes but also achieve better accuracy comparing to individual components. In this experiment the threshold τ_1 and τ_2 are set to 3 and 0.5 respectively. However, the result might seem a bit bias since the evaluation is performed by only one subject.

4.2 Subjective Evaluation

A subjective evaluation is further conducted to examine and emphasize the accuracy and effectiveness of the proposed method – LPO from a variety of users. As aforementioned, there were 263 keywords generated using LPO model which is inconvenient and time-consuming for users to evaluate the result. Therefore, we categorize purposes and ask users to enumerate important features for each purpose category. We manually defined nine categories, while this categorization could be done automatically in the future work by mining both taxonomic and non-taxonomic relations as in [14] (Table 6).

Questionnaire Setup. We created 10 questions in total [15]. Question #1 asks about the respondent's experiences in photography skill. Question #2 – Question #10 survey respondent's knowledge or experiences for which relevant camera features they are using for each purpose. Relevant camera features per purpose is defined by selecting camera features receiving 50% votes from all respondents.

Table 6. Sample of purpose categories and individual keywords

Purposes	Keywords
Macros	flower, insects, foods, ...
Moving targets	animals, kids, freezing, motion, action shots, sport events, soccer, ...
Traveling	trips, touring, backpack, adventures, outings, ...
Low light condition	dimmer environments, backlit daylight, dark room, ...
Bright condition	sunlight, daylight, outdoor, portraits at outdoor, ...
...	...

Result. We received 31 respondents in which the majority of respondents are beginner users (less than 1 year) and professional users (more than 5 years). The responses from them could possibly be viewed as subjective evaluations with different opinions and diversities. Since we asked people to rank the camera features for different purposes, we could obtain the ground truth of relationship between user purposes and features and able to calculate the accuracy of each method as shown in Tables 7 and 8. Though the precision of pattern-based models and LPO is the same, but the number of individual keywords extracted by the LPO is higher based on evaluation in Sect. 4.1. Therefore, we could conclude that our proposed model could extract more good purposes (better recall) than the pattern-based and ontology-based separately with the best accuracy of F-score 0.67.

Table 7. Category-based precision and recall of proposed model – LPO

Purpose category	Ground truth	Proposed model	P	R
Macros/Closeups	Focus, IS, ISO	Focus, IS	1	0.67
Landscape/Scenery	ISO, Focus, IS	Focus, Shutter, IS	0.67	0.67
Moving targets	Shutter, ISO, Focus, IS	Focus, Shutter, IS	1	0.75
Portraits	Focus, IS, ISO	Focus, IS	1	0.67
Viewing and framing	Screen	Screen	1	1
Uploading, editing, and printing	N/A	Pixel	0	N/A
Low light condition	ISO, Flash	ISO, Flash	1	0.5
Bright condition	ISO, Shutter	ISO, Flash	0.5	0.5
Traveling	Shutter, Focus, IS	Weight	0	0
Average			0.69	0.66

Table 8. Precision, recall, and F-score of each method (Category-based)

Methods	P	R	F-score
Lexicon	0.69	0.54	0.61
POS	0.69	0.61	0.65
LP	0.69	0.61	0.65
O	0.11	0.06	0.08
LPO (Proposed Model)	0.69	0.66	0.67

5 Related Work

This paper deals with various fields of studies, including feature-based ranking system, mining semantic relations using bootstrapping, and ontology learning. We will take a look at some of the previous works related to each field of study.

Feature-Based Ranking System. Mining of opinions from reviews has become a popular area of research. Users often express their usage experiences in reviews. Regarding the feature extraction techniques, it could be classified into supervised and unsupervised approach. An example of supervised techniques, Wong and Lam [16] employ the hidden Markov model and conditional random fields as the underlying learning method for extracting product features from auction websites. In contrast, product features can also be extracted automatically using unsupervised technique presented by Hu and Liu [17]. They made assumptions that product features must be noun or noun phrases. This utilizes the association rule mining algorithm to discover all frequent item sets within a target set of reviews. Regarding the problem of feature-based summarization, Kangale et al. [3] provided a solution which could extract user opinions from reviews and generate rating as well as review summary of each product feature. Zhang et al. [1] proposes a methodology to rank products based on their features using online reviews. First, they manually define a set of product features that are of interest to the customers. Next, they detect subjective and comparative sentences in reviews using predefined structural patterns. Using these sentences, they construct a feature-specific product graph and apply page-rank like algorithm to rank products based on the product features. Our work differs from theirs in the following aspects: (1) we assume that most of the consumers do not have much knowledge about product features and attributes and thus, they could not prioritize which product feature is important or relevant and (2) instead of generating a review summary on each product feature or product as a whole, we are interested only on extracting the semantic relationship between product features and user purposes.

Mining Semantic Relations with Bootstrapping. Most common approach of bootstrapping to extract the semantic relations is pattern-based approach. Agichtein and Gravano [9] proposed “Snowball” algorithm to identify and extract structured relations between named entities from unstructured text. The system is given with some initial seeds then searches sentences in which terms in a seed tuple occur closely to each other. It then analyzes the connecting text and surrounding context and generates an extraction pattern. Pantel et al. [13] proposed an approach based on an edit-distance technique to learn lexicon-POS patterns and could obtain both good performance and efficiency. Later Espresso [18] has been proposed by the same author to infer patterns for harvesting binary semantic relations (is-a, part-of, succession, reaction, and production). It also describes refining techniques to deal with wide variety of relations by measuring the strength of association between patterns and seeds using pointwise mutual information. In our work, we propose an additional filtering approach for accepting new seeds by measuring the word semantic similarity to avoid the semantic drift.

Ontology Learning. Because we have introduced an ontology learning approach to generate more seed instances in conventional bootstrapping, we will take a look at some studies related to ontology. Hearst [11] pioneered using patterns to extract hyponym (is-a) relations. Manually building a few lexicon-syntactic patterns, Hearst sketched a bootstrapping algorithm to learn more patterns from instances, which has served as the model for most subsequent pattern-based algorithms. We have adopted this approach in our study by extracting hyponym relation of generated seeds using some predefined lexicon-syntactic patterns listed in Table 3. Chen et al. [14] proposed an ontology learning framework to extract customer needs of digital cameras. To detect the ontology relation, the authors detect taxonomic and non-taxonomic relations from the customer reviews. Taxonomic relations are extracted using several methods including string matching, lexicon-syntactic patterns, and WordNet taxonomic relations. They proposed a word property-based method for extracting non-taxonomic relations by measuring the support and confidence of the co-occurrence noun phrases from the corpus.

6 Conclusion and Future Work

In this paper, we proposed a methodology for extracting semantic relations between product features and purposes. This work is necessarily important in order to realize the proposed framework of purpose-oriented recommendation. The output of the framework could help educating novice users to comprehend the product features which should be taken into account for their specific needs or purposes. The result of experiments indicates that the combination of pattern-based and ontology-based approach could extract number of good purposes and obtain a good result of precision and recall.

For the future research, we are considering to quantify the optimal setting or requirements of product features for achieving a specific purpose. For instance, if the user is interested in buying a digital camera for shooting their busy kids, then it is desired to not only rank the relevant features but also the optimal setting configuration (shutter speed of 1/500 s, 5 fps burst rate, ISO 1600, continuous AFC etc. is at least required to capture sharp image). Such quantification not only tells users which camera they should purchase considering the important features but also educates them to set the right setting for achieving their goal when they use the product.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number 15K00423 and the Kayamori Foundation of Informational Science Advancement.

References

1. Zhang, K., Narayanan, R., Choudhary, A.: Voice of the customers: Mining online customer reviews for product feature-based ranking. In: Proceeding of 3rd Workshop on Online Social Networks (WOSN), pp. 1–9 (2010)
2. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceeding of 19th National Conference on Artificial Intelligence, pp. 755–760 (2004)

3. Kangale, A., Kumar, S.K., Naeem, M.A., Williams, M., Tiwari, M.K.: Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary. *Int. J. Syst. Sci.* **47**, 3272–3286 (2016)
4. Kamal, A.: Review mining for feature based opinion summarization and visualization. *arXiv Preprint [arXiv:1504.03068](https://arxiv.org/abs/1504.03068)* (2015)
5. Uchida, S., Yamamoto, T., Kato, M.P., Ohshima, H., Tanaka, K.: Object search by experience attributes. *DBSJ Japanese J.* **14-J**(7), 1–7 (2016)
6. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: *Proceeding of EMNLP 2009*, pp. 248–256 (2009)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: *Proceeding of Workshop at International Conference on Learning Representations (ICLR)* (2013)
8. Loper, E., Bird, S.: NLTK: the natural language toolkit. In: *Proceeding of ETMTNLP 2002*, vol. 1, pp. 63–70 (2002)
9. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: *Proceeding of 5th ACM Conference on Digital libraries*, pp. 85–94 (2000)
10. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: *Proceeding of EMNLP 2002*, pp. 214–221 (2002)
11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: *Proceeding of 14th Conference on Computational Linguistics*, pp. 539–545 (1992)
12. McAuley, J., Pandey, R., Leskovec, J.: Inferring networks of substitutable and complementary products. In: *Proceeding 21th KDD*, pp. 785–794
13. Pantel, P., Ravichandran, D., Hovy, E.: Towards terascale knowledge acquisition. In: *Proceeding of 20th International Conference on Computational Linguistics*, pp. 771–777 (2004)
14. Chen, X., Chen, C.-H., Fai Leong, K., Jiang, X.: An ontology learning system for customer needs representation in product development. *Int. J. Adv. Manuf. Technol.* **67**, 441–453 (2013)
15. Online Survey about Digital Cameras, <https://www.surveymonkey.com/r/RMW9C7T>
16. Wong, T.-L., Lam, W.: Learning to extract and summarize hot item features from multiple auction web sites. *Knowl. Inf. Syst.* **14**, 143–160 (2008)
17. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: *Proceeding of 10th KDD*, pp. 168–177 (2004)
18. Pantel, P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations patrick. In: *Proceeding of Joint Conference of 21st COLING and 44th ACL*, pp. 113–120 (2006)

Data Mining and Big Data

Second International Conference, DMBD 2017,

Fukuoka, Japan, July 27 – August 1, 2017, Proceedings

Tan, Y.; Takagi, H.; Shi, Y. (Eds.)

2017, XV, 546 p. 210 illus., Softcover

ISBN: 978-3-319-61844-9