Mining Attribute-Specific Ratings from Reviews of Cosmetic Products

Yuuki Matsunami, Mayumi Ueda, Shinsuke Nakajima, Takeru Hashikami, John O'Donovan, and Byungkyu Kang

1 Introduction

In recent years, many online sellers of cosmetic products have added support for user-provided reviews. These are very helpful for consumers to decide whether to buy a commercial product, and they have been shown to have a significant impact on conversion rates. In particular, consumers make careful choices about cosmetics since unsuitable products frequently cause skin irritations. "@cosme" [1] is a cosmetics review site that is very popular among Japanese young women. While the site can be helpful in decision making, it is not easy task to find truly suitable cosmetics because of the lack of explanation and granularity in user provided ratings of products on the site. As an example, there is no guarantee that a cosmetics item, that one contributor mentioned as being good for dry skin, is always suitable for people who have dry skin. Since the compatibilities between skin and cosmetics items differ from one user to another, we believe it is important to identify and cluster users who share common preferences for cosmetic products and to share

Y. Matsunami (✉) • S. Nakajima
Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, 603-8555, Kyoto, Japan
e-mail: g1245108@cc.kyoto-su.ac.jp; nakajima@cse.kyoto-su.ac.jp

M. Ueda
University of Marketing and Distribution Sciences, 3-1 Gakuen-Nishimachi, 651-2188, Nishi-ku, Kobe, Hyogo, Japan
e-mail: Mayumi_Ueda@red.umds.ac.jp

T. Hashikami
istyle Inc., 1-12-32 Akasaka Minato-ku, 107-6034, Tokyo, Japan
e-mail: hashikamit@istyle.co.jp

J. O'Donovan • B. Kang
University of California, Santa Barbara, 93106, Santa Barbara, CA, USA
e-mail: jod@cs.ucsb.edu; bkang@cs.ucsb.edu

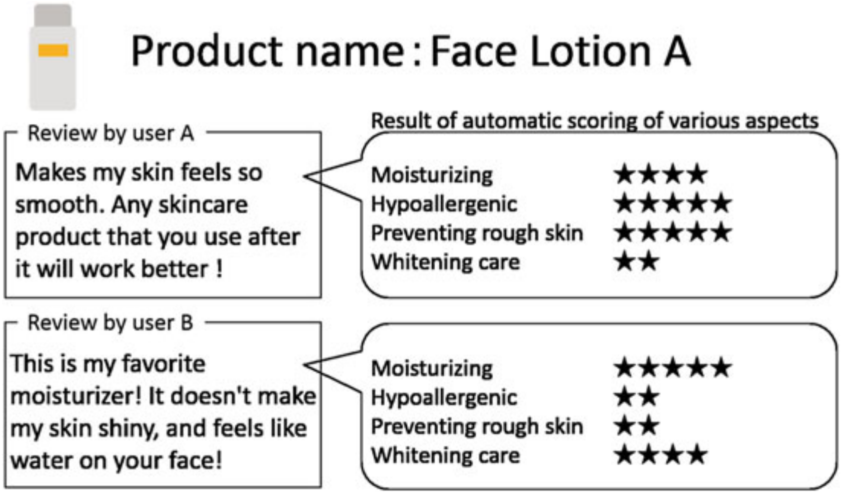
S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-3950-8_8

reviews among those niche communities. To study the proposed approach, we design and evaluate a collaborative recommender system for cosmetic products, which incorporates opinions of similar-minded users and automatically scores fine grained aspects of product reviews.

In order to develop such a review recommender system, we have to analyze review text to understand feedback on reviewers' experiences of cosmetic items. Actually, there is a score (as # of stars) of each review text on the conventional cosmetic review sites. However, it is typically an overall score for an item, meaning that it can mask the experiences of reviewers with different aspects of the item. For example, there are attributes such as "moisturizing effect", "whitening effect", "exfoliation care effect", "Hypoallergenic effect", and "Aging care effect", contained in reviews for "face lotion". Thus, we need a scoring method of such various aspects of cosmetic item review texts to understand feedback on reviewers' experiences at this finer grained, attribute-specific level.

Hence, the purpose of our study is to propose a method for automatic scoring of various aspects of cosmetic item review texts based on evaluation expression dictionary. The method can realize an automatic scoring of various aspects of cosmetic item reviews even if no scores are explicitly mentioned (see Fig. 1). In this paper, we construct an evaluation expression dictionary for "face lotion", which has different feedback with each person, as a first step. Moreover, we discuss the suitability of our proposed method based on an evaluation experiment for the automatic scoring method. This paper is revised version of the conference paper that we gave a presentation at IMECS2016 [2].

The remainder of this paper is organized as follows. The related work is given in Sect. 2. Then Sect. 3 describes the method for automatic scoring of various aspects of cosmetic item review texts based on evaluation expression dictionary. Discussions about the utility of our proposed method based on an evaluation experiment for the automatic scoring method are given in Sect. 4. We conclude the paper in Sect. 5 with a discussion of key results and avenues for future work.


Product name : Face Lotion A
Review by user A
Result of automatic scoring of various aspects
Review by user B

Makes my skin feels so smooth. Any skincare product that you use after it will work better !

Moisturizing ★★★★
Hypoallergenic ★★★★★
Preventing rough skin ★★★★★
Whitening care ★★

This is my favorite moisturizer! It doesn't make my skin shiny, and feels like water on your face!

Moisturizing ★★★★★
Hypoallergenic ★★
Preventing rough skin ★★
Whitening care ★★★★

Fig. 1 Example of automatic scoring of various aspects of cosmetic item reviews

2 Related Work

There are many websites that support user-provided reviews on products. For example, Amazon.com [3] and Priceprice.com [4] are popular shopping sites, and these sites provide mechanisms for their customers to leave reviews on products they have purchased. And "Tabelog" is also a popular website in Japan. This website does not sell physical products. Instead, it provides restaurant information and reviews. In addition to the algorithmic aspects, researchers have recently focused on the presentation aspects of review data [5]. Furthermore, in recent years, "@cosme" has become very popular among Japanese young women. This website is a portal site for beauty and cosmetic items, and it provides various information, such as reviews and shopping information for cosmetic items. According to the report by the istyle Inc. that is a parent company of this system, as of November 2015, the number of monthly page views was 280 million, the number of members was 3.5 million, and the total number of reviews was 1200 million [6]. From this report, it is clear that many women exchange information about beauty and cosmetics through the @cosme service. @cosme provide information about cosmetic items of various cosmetic brands. Hence, users can compare cosmetic items through the various cosmetic brands. Reviews are composed of review text, scores, tags about effects, etc. Furthermore, the system has profile data that includes information about age and skin type, provided by the users when they enrolled as a member. Therefore, users who want to browse the reviews can search the reviews according to their own purposes, for example, reviews sorted by the scores or focused on one effect.

Along with the popularization of these review services, several studies about analysis of reviews have been conducted in the past. For example, O'Donovan et al. evaluated their AuctionRules algorithm – a dictionary-based scoring mechanism for eBay reviews of Egyptian antiques [7]. They showed that the approach was scalable and particularly that a small amount of domain knowledge can greatly improve prediction accuracy compared against traditional instance-based learning approach. In our previous study, we analyze reviews of the cosmetic items [8]. In order to determine if the review is positive review or negative review, we make dictionaries for the Japanese language morphological analysis, which composed of positive expression and negative expression of cosmetic items. This previous research is aimed to develop the system to provide the reviews that take account of the user's profile, then, that system tries to retrieve information from blogs and SNS, and attempts to merge the information to the same format. The final goal of our current study is to develop a method for automatic scoring of review texts, according to various aspects of cosmetic items.

Nihongi et al. propose a method for extracting the evaluation expression from the review texts, and they develop the product retrieval system using evaluation expressions [9]. Our research focuses on the analysis of the review for cosmetic items, and we are aimed at finding similar users based on preferences and feelings in order to recommend truly useful reviews.

Titov et al. propose a statistical model for sentiment summarization [10]. This model is a joint model of text and aspect ratings. In order to discover the corresponding topics, this model uses aspect ratings. Therefore, this model is able to extract textual evidence from reviews without the need for annotated data.

As stated above, there are several studies that attempt to analyze reviews. However, there has been no study that tried to develop a method for automatic scoring of review texts, according to fine-grained aspects of cosmetic items. In the following section, we outline our novel approach to this problem.

3 Approach

In this section, we describe a method for automatic scoring of various aspects of cosmetic item review texts based on an evaluation expression dictionary. First, we describe the brief overview of our proposed method in Sect. 3.1. Section 3.2 explains how to construct the evaluation expression dictionary. The method for automatic scoring of various aspects of cosmetic item review texts is then given in Sect. 3.3.

3.1 Overview of Proposed Method

The purpose of this paper is to propose a method for automatic scoring of various aspects of cosmetic item review texts based on an evaluation expression dictionary. Furthermore, our final goal is to develop a cosmetic item review recommender system which can recommend truly useful reviews for a target user. It operates in a similar manner to collaborative filtering, by using a set of similar users who have common both preferences and feedbacks on their experiences of the cosmetic items.

In order to make a significance of our study clear, Fig. 2 shows a conceptual diagram of the cosmetic item review recommender system, which is our final goal. In Fig. 2, numbers in blue written as (1)–(4) are corresponding to the procedure of cosmetic review automatic scoring process, and Roman alphabets in red written as (a)–(e) are corresponding to the procedure of review recommendation process. More detailed procedures are shown below:

Automatic Scoring

1. Construct the evaluation expression dictionary which includes pairs of evaluation expression and its score by analyzing reviews sampled from non-scored DB.
2. Pick up reviews from non-scored DB to score them.
3. Automatically score reviews picked up in step (2) based on the evaluation expression dictionary constructed in step (1).
4. Put reviews scored in step (3) into scored review DB.

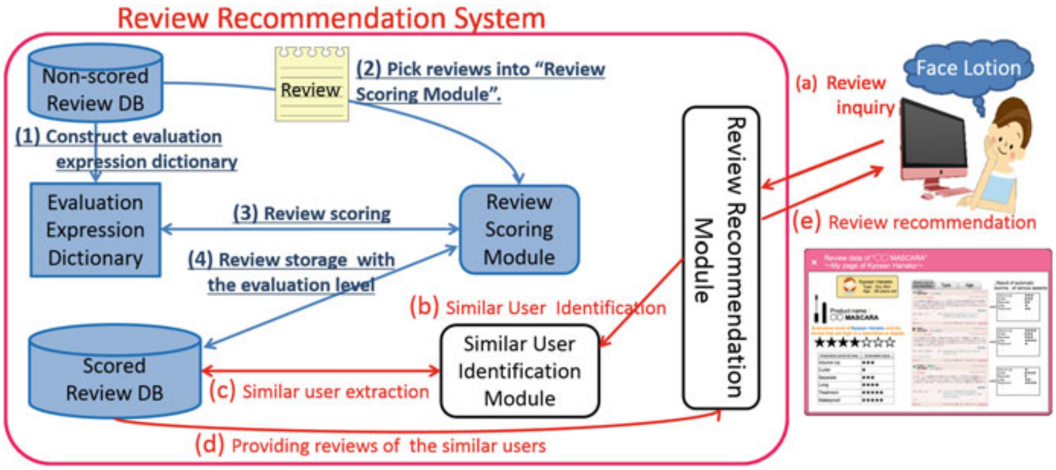

Review Recommendation System
Non-scored Review DB
Review
(2) Pick reviews into "Review Scoring Module".
(1) Construct evaluation expression dictionary
Evaluation Expression Dictionary
(3) Review scoring
Review Scoring Module
(4) Review storage with the evaluation level
(b) Similar User Identification
Scored Review DB
(c) Similar user extraction
Similar User Identification Module
Review Recommendation Module
(d) Providing reviews of the similar users
(a) Review inquiry
Face Lotion
(e) Review recommendation
MASCARA

Fig. 2 Conceptual diagram of the cosmetic item review recommender system

Review Recommendation Process

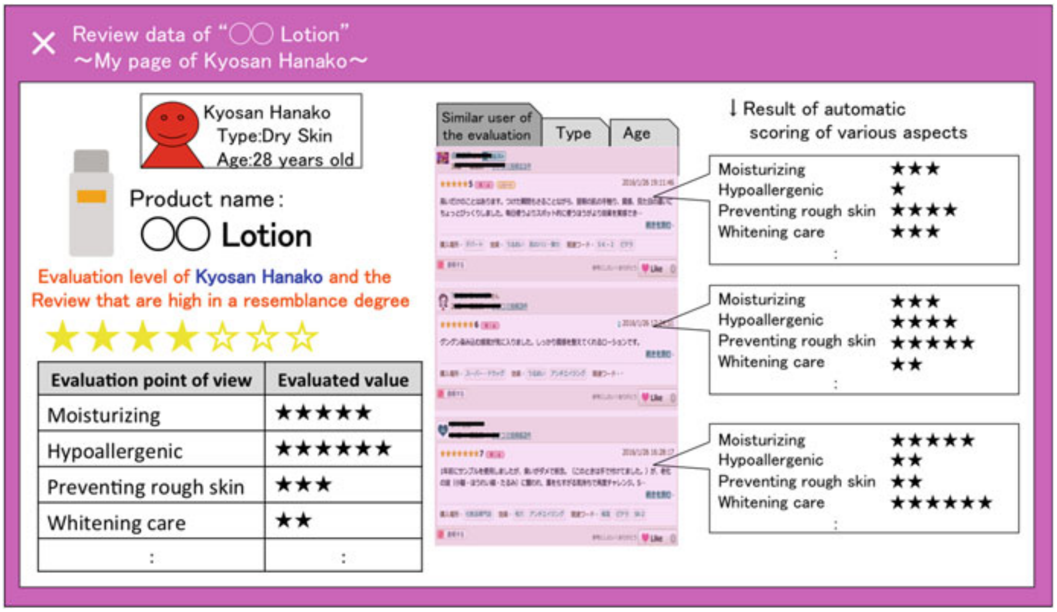
(a) User provides the name of a cosmetic item that she is interested in.
(b) System Refers to "similar user extraction module" in order to extract similar users to the target user of step (a).
(c) "Similar user extraction module" obtains the information about reviews and reviewers, and identies similar users to the target user.
(d) Provide reviews of the similar users identified in step (c) to "Review recommendation module".
(e) System recommends suitable reviews to the target user.

This paper focuses on development of the dictionary-based approach. Developing the live review recommendation method is part of our future work.

Figure 3 shows an example of user interface of the cosmetic item review recommender system. We believe that users can browse truly suitable reviews of a target cosmetic item and also they can easily choose reviewers group such as "reviewers having similar evaluation tastes", "reviewers having a similar skin type" and "reviewers of the same age group" by clicking the tab. Moreover, the user interface can provide not only reviews themselves but also their scores for the various aspects against the target cosmetic item, so that users can understand what kind of feedback on reviewers' experiences of cosmetic items without difficulty.

3.2 Constructing the Evaluation Expression Dictionary

We describe how to construct the evaluation expression dictionary which has pairs of evaluation expressions and the scores against cosmetic items in this section.


Review data of "〇〇 Lotion"
～My page of Kyosan Hanako～
Kyosan Hanako
Type:Dry Skin
Age:28 years old
Product name:
〇〇 Lotion
Evaluation level of Kyosan Hanako and the Review that are high in a resemblance degree
★★★★☆☆☆
Similar user of the evaluation
Type
Age
↓Result of automatic scoring of various aspects

Evaluation point of view	Evaluated value
Moisturizing	★★★★★
Hypoallergenic	★★★★★★
Preventing rough skin	★★★
Whitening care	★★
:	:

Moisturizing	★★★
Hypoallergenic	★
Preventing rough skin	★★★★
Whitening care	★★★
:	

Moisturizing	★★★
Hypoallergenic	★★★★
Preventing rough skin	★★★★★
Whitening care	★★
:	

Moisturizing	★★★★★
Hypoallergenic	★★
Preventing rough skin	★★
Whitening care	★★★★★★
:	

Fig. 3 An example of user interface of the cosmetic item review recommender system

Cosmetic item reviews can come from different people with widely varying expressions. Thus, we try to construct the dictionary by extracting and registering evaluation expressions from real and diverse review data. We gather review data to construct the dictionary from @cosme [1] which is the representative cosmetic review website in Japan. In particular, we extract both frequently-appearing expressions in good evaluations and bad evaluations for each cosmetic item, and register these evaluation expressions into the dictionary.

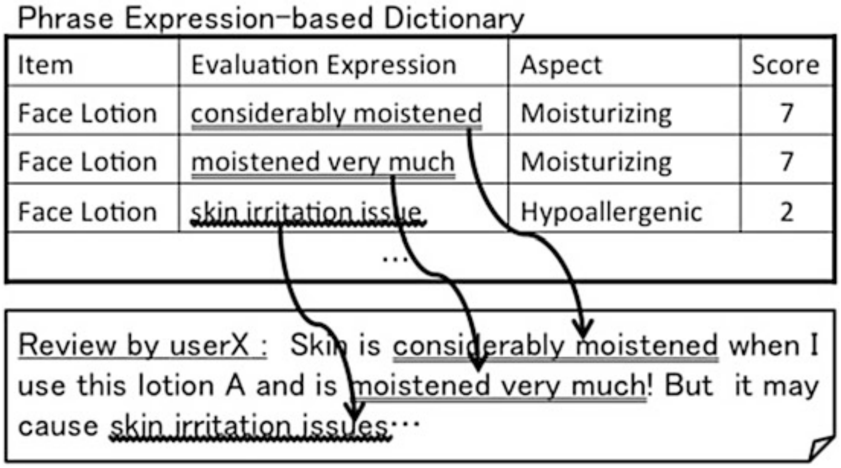
There are many kinds of cosmetic items. As a first step, we try to construct evaluation expression dictionary of "face lotion" in this paper. The reason why we focus on "face lotion" is that "face lotion" is used by a lot of people. In addition, there are various evaluations against even one product of "face lotion" due to differences of users' skin types.

3.2.1 Phrase Expression-Based Dictionary

In order to construct the phrase expression-based dictionary, we gather 80 reviews for "face lotion", and manually extract 1,893 characteristic evaluation phrases from them. Next, two evaluators, who are both female students in their 20's, give a score against each evaluation expression phrase manually, and we set an average of the scores as the final score of the expression phrase. There are widely various expressions in review texts because they are based on free description in natural languages. Therefore, we categorize gathered evaluation expressions into 39 groups which correspond to detailed Categories in Table 1 by consulting the effect-tags in @cosme. Figure 4 shows the data format of the phrase expression-based dictionary.

Table 1 Categories of evaluation expression against “face lotion”

Rough categories	Medium categories	Detailed categories
Cost performance	Cost performance	Cost
Moisturizing/penetration	Moisturizing	Keeping moisture, moist, water, dry/dry skin
		Moisturizing, fresh and young
	Penetration	Skin familiarity, penetration, suction
	Tenseness and elasticity	Elasticity, springy, stick to
	Tightening the skin	Tightening the skin
Whitening care/UV	Whitening care	Whitening care, dullness, transparency
	UV care	UV care
Exfoliation & pores care/cleansing effect	Exfoliation care	Exfoliation
	Pores care	Pores
	Cleansing effect	Cleansing
Refreshing feeling/preventing sebum shine	Refreshing feeling	Refreshing condition, refreshing feeling
	Preventing sebum shine	Tacky, oil, shine
Refreshing ↔ thickening	Refreshing ↔ thickening	Refreshing texture, thickening, sense of use
Hypoallergenic	Hypoallergenic	Sensitive skin, stimulation
	Organic	Organic
Preventing rough skin	Preventing rough skin	Skin roughness, trouble
	Acne care	Acne care
Aging care	Anti-aging	Anti-aging, beauty ingredient
Fragrance	Fragrance	Fragrance, healing



Phrase Expression-based Dictionary

Item	Evaluation Expression	Aspect	Score
Face Lotion	considerably moistened	Moisturizing	7
Face Lotion	moistened very much	Moisturizing	7
Face Lotion	skin irritation issue	Hypoallergenic	2
…			

Review by userX : Skin is considerably moistened when I use this lotion A and is moistened very much! But it may cause skin irritation issues…

Fig. 4 Review scoring using phrase expression-based dictionary

The procedure of automatic scoring method based on the phrase expression-based dictionary is as follows:

At first, the method gathers non-scored reviews, and identify evaluation expressions existing in these reviews. Secondly, it gives a score to each evaluation expression based on the dictionary if there is same evaluation expression in the dictionary.

For example in Fig. 4, the review text includes phrases as “considerably moistened” and “moistened very much” related to an aspect of “Moisturizing”, so that the method give a score “7” as an average of their scores based on the phrase expression-based dictionary. Moreover, it includes phrases as “skin irritation issues” related to an aspect of “Hypoallergenic”. Thus, the method can give a score “2” based on the dictionary.

Next, we examine a scoring test for non-scored review data based on the constructed dictionary, and compare the result with ground truth data in order to evaluate the effectiveness of the phrase expression-based dictionary. The ground truth data is provided based on not the dictionary but manual detection. We use 16 non-scored reviews and compare numbers of evaluation expressions that are scored by each method in this test.

As a result of the scoring test, a number of evaluation expressions detected manually is 101, whereas a number of evaluation expressions scored based on the phrase expression-based dictionary is 5. That is about only 5% of ground truth data. The reason for this result may be that it is very difficult to construct a phrase expression-based dictionary that can cover various evaluation expression phrases in a large amount of reviews. Because there are many kinds of phrasal expressions to express a single meaning. Therefore, we think that it is necessary to construct not a phrase expression-based dictionary but another dictionary which can cover more evaluation expressions.

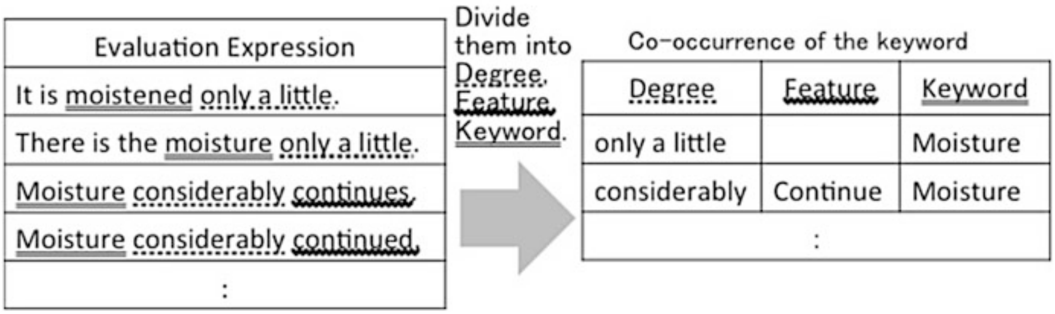
3.2.2 Co-occurrence Keyword-Based Dictionary

As mentioned in previous section, it is difficult for a phrase expression-based dictionary to cover most evaluation expressions in a lot of reviews. Thus, we try to construct another dictionary using co-occurrence keyword-based evaluation expressions in order to cover wider scope of evaluation expressions.

Figure 5 describes conceptual diagram of constructing the co-occurrence keyword-based dictionary. The procedure of constructing the dictionary is as follows:

1. Analyze phrasal evaluation expressions extracted from reviews.
2. Divide the phrasal expressions into aspect keywords, feature words and degree words.
3. Construct the dictionary by assembling their co-occurrence relations and the evaluation scores.

Figure 6 shows differences of detecting evaluation expression between phrase expression-base dictionary and co-occurrence keyword-base dictionary. As shown

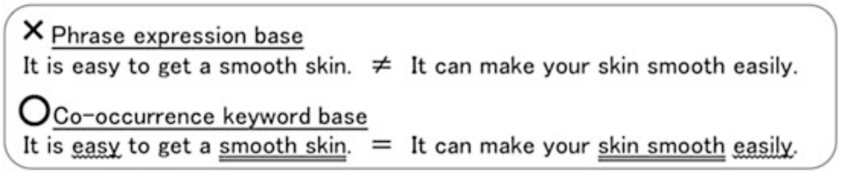


Evaluation Expression
It is moistened only a little.
There is the moisture only a little.
Moisture considerably continues.
Moisture considerably continued.
:

Co-occurrence of the keyword

Degree	Feature	Keyword
only a little		Moisture
considerably	Continue	Moisture
:		

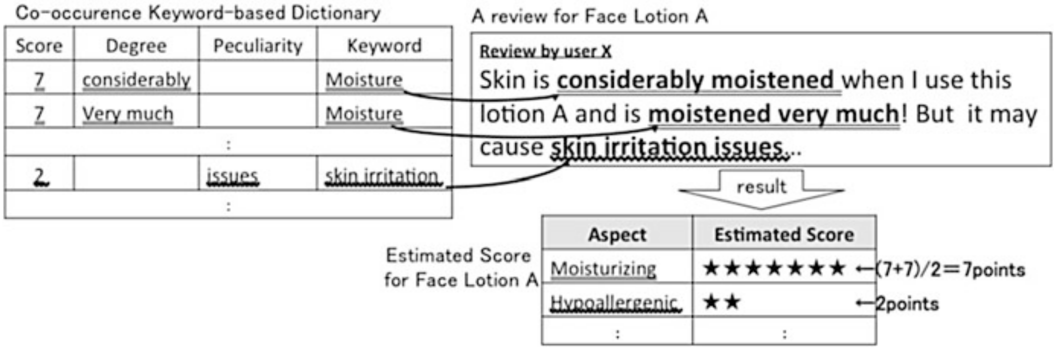
Fig. 5 Constructing the co-occurrence keyword-based dictionary



× Phrase expression base
It is easy to get a smooth skin. ≠ It can make your skin smooth easily.

O Co-occurrence keyword base
It is easy to get a smooth skin. = It can make your skin smooth easily.

Fig. 6 Differences of detecting evaluation expression between phrase expression-base and co-occurrence keyword-base



Co-occurence Keyword-based Dictionary

Score	Degree	Peculiarity	Keyword
7	considerably		Moisture
7	Very much		Moisture
:			
2		issues	skin irritation
:			

A review for Face Lotion A

Review by user X
Skin is considerably moistened when I use this lotion A and is moistened very much! But it may cause skin irritation issues...

Estimated Score for Face Lotion A

Aspect	Estimated Score
Moisturizing	★★★★★★★ ←(7+7)/2=7points
Hypoallergenic	★★ ←2points
:	:

Fig. 7 Automatic scoring based on the co-occurrence keyword-based dictionary

in Fig. 6, a phrase “It is easy to get a smooth skin” and another phrase “It can make your skin smooth easily” are semantically nearly identical but are different as a phrase. Hence, it may be possible to detect more evaluation expressions based on the co-occurrence keyword-base dictionary than based on the phrase-based dictionary.

3.3 Automatic Scoring Based on Evaluation Expression Dictionary

The procedure of automatic scoring against non-scored reviews based on the evaluation expression dictionary is shown below (see Fig. 7):

1. System examines a morphological analysis against non-scored review data, and investigates existence or non-existence of aspect keywords as evaluation expression for cosmetic items in the review data.
2. If an aspect keyword exist in it, system checks presence or absence of co-occurrence feature words and degree words co-occurring with the aspect keyword.
3. System make an inquiry on the dictionary to get the score of the evaluation expression based on an aspect keyword, a feature word and a degree word.
4. System achieve "automatic scoring of various aspects of review texts" by aggregating such scores for each aspect in a review.

4 Experimental Evaluation of Automatic Scoring Using Real Review Data

We examine an experimental evaluation of the automatic scoring method using real review data in order to verify the effectiveness of our proposed method. As a first step, we analyze 5,000 reviews randomly extracted from review data for "face lotion" posted at @cosme, for understanding characteristics of the data.

Figure 8 shows a number of reviews for each star (score), and Fig. 9 describes an average number of tags for each star (score) in the 5,000 reviews. Reviewers can

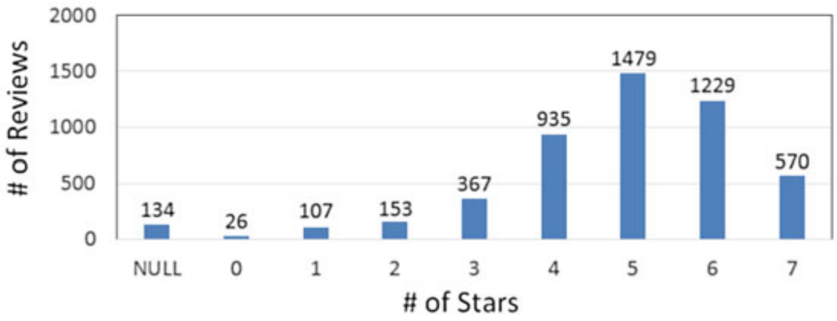


Fig. 8 Number of reviews for each star

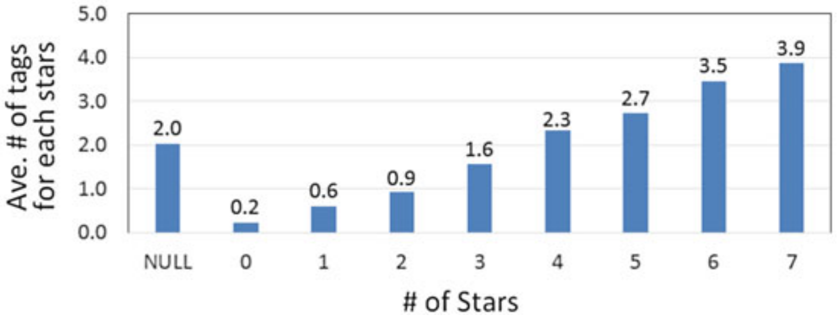


Fig. 9 Average number of tags for each star

describe stars from 0 to 7 as score and tags to describe the effectiveness of cosmetics items at the @cosme website.

According to Fig. 8, the average number of stars is 4.94 and the distribution of the data looks balanced. According to Fig. 9, the average number of tags is 2.75 and we can see that reviewers who give a good score tend to provide more tags.

4.1 Procedure of Experimental Evaluation

In this experiment, we use 10 review data for "face lotion" randomly selected from 5,000 reviews as described above, and then compare results by the following methods:

- Manual scoring method without the Dictionary (as ground truth data).
- Automatic scoring method based on the evaluation expression dictionary (proposed method).

In the case of the manual scoring, evaluators actually read review texts and score them between 0 to 7 stars for 10 aspects of "face lotion" set in advance. The evaluators are 30 people. They are 20's to 50's females.

In the case of automatic scoring, the method scores review texts between 0 to 7 stars for the 10 aspects based on co-occurrence keyword-based dictionary.

The 10 aspects for "face lotion" set for the experiment in advance are shown below:

- Cost performance
- Moisturizing
- Whitening care
- Refreshing feeling/Preventing sebum shine
- Refreshing↔Thickening
- Hypoallergenic
- Preventing rough skin
- Aging care
- Fragrance

4.2 Result of the Experiment

Figure 10 describes results of review scoring against 10 reviews based on co-occurrence keyword-based dictionary and manual scoring by evaluators.

The contents of 10 reviews are different from each other, so that the detected aspects are different. The average score (# of stars) of all aspects by manual scoring is 4.77, and the average score by our proposed method is 4.52. The mean absolute error (MAE) is 0.70.

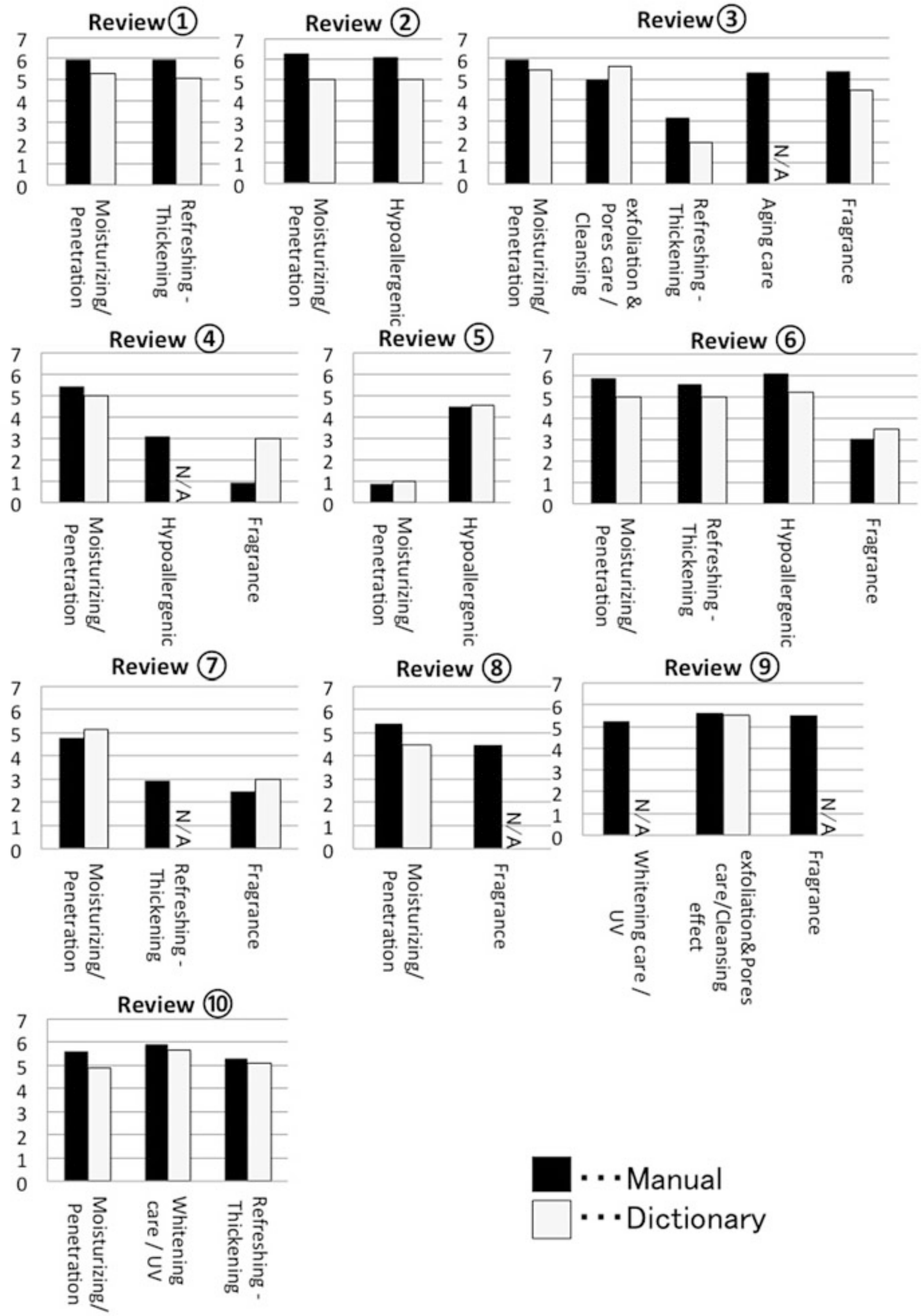


Fig. 10 Result of review scoring based on co-occurrence keyword-based dictionary

Scores of the manual scoring tend to a little higher than scores of prosed scoring method. However, the range of the score is from 0 to 7 and MAE is 0.70, so that we may say that the results of automatic scoring by proposed method are quite close to the results of manual scoring as ground truth data. Total number of detected aspects are 29 aspects by manual scoring (ground truth) and 23 aspects by automatic scoring (proposed method). Therefore, the achievement rate of our proposed method against manual scoring is about 79%. The achievement rate of phrase-based dictionary is about 5% as shown in Sect. 3.2.1. Thus, the result of our proposed method based on co-occurrence keyword-based dictionary indicates high potential for detecting aspects of cosmetic items.

There are several “N/A” in Fig. 10 by automatic scoring method. However, there is room for improving the result of aspect detection for reviews by updating the dictionary. In future work we will try to analyze larger number of reviews, and then improve and tune the dictionary. Moreover, we will develop a review recommendation system for cosmetic items to further evaluate our novel scoring method.

5 Conclusions

In this paper, we presented a method for automatic scoring of various aspects of cosmetic item review texts based on an evaluation expression dictionary. In order to realize our proposed method, we constructed two types of evaluation expression dictionaries by extracting and registering evaluation expressions from real review data. Firstly, we constructed a phrase expression-based dictionary. However, it is difficult to cover most evaluation expression in a lot of reviews. Therefore, secondly, we constructed another dictionary using co-occurrence keyword-based evaluation expressions in order to cover the wide scope of evaluation expressions. In order to verify the accuracy of our proposed method, we conducted a simple experiment for the automatic scoring method. We will improve the dictionary to cover more evaluation expressions, in future work. A future direction of this study will also be to develop a cosmetic item review recommender system which can recommend truly useful reviews for a target user, by leveraging the approach we describe here.

Acknowledgements This work was supported in part by the MEXT Grant-in Aid for Scientific Research(C)(#16K00425, #26330351).

References

1. @cosme, http://www.cosme.net/
2. Y. Matsunami, M. Ueda, S. Nakajima, T. Hashikami, S. Iwasaki, J. O’Donovan, B. Kang, Explaining item ratings in cosmetic product reviews, in *Proceedings of the International Multiconference of Engineers and Computer Scientists 2016*, Hong Kong, 16–18 Mar 2016. Lecture Notes in Engineering and Computer Science, pp. 392–397

3. Amazon.com, http://www.amazon.com/
4. Priceprice.com, http://ph.priceprice.com/
5. B. Kang, N. Tintarev, J. O'Donovan, Inspection mechanisms for community-based content discovery in microblogs, in *IntRS'15 Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, Vienna (2015). http://recex.ist.tugraz.at/intrs2015/ at ACM Recommender Systems 2015
6. The site data of @cosme (Nov 2015), istyle Inc. http://www.istyle.co.jp/business/uploads/\sitedata.pdf (in Japanese)
7. J. O'Donovan, V. Evrim, P. Nixon, B. Smyth, Extracting and visualizing trust relationships from online auction feedback comments, in *International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad (2007)
8. Y. Hamaoka, M. Ueda, S. Nakajima, Extraction of evaluation aspects for each cosmetics item to develop the reputation portal site, in *IEICE WI2-2012-15* (2012, in Japanese), pp. 45–46
9. T. Nihongi, K. Sumita, Analysis and retrieval of the word-of-mouth estimation by structurizing sentences, in *Proceeding of the Interaction 2002* (2012, in Japanese), pp. 175–176
10. I. Titov, R. McDonald, A joint model of text and aspect ratings for sentiment summarization, in *46th Meeting of Association for Computational Linguistics(ACL-08)*, Columbus (2008), pp. 308–316