The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation

Atsuho Nakayama

Abstract This study attempted to detect trending topics and temporal variation in web communication topics regarding new products among consumers using social media. This was done by classifying words into clusters based on their co-occurrence. We collected Twitter entries about new products based on their specific expressions of sentiment or interest. Because of the desire to identify market trends, the analysis of consumer tweet data has received much attention. To construct appropriate words, we used a complementary similarity measure, a classification method that is widely applied in character recognition. We classified the words extracted from Twitter data using non-negative matrix factorization as a dimensionality reduction model. To help interpret the results, we proposed a visualization method for text classification using a multidimensional scaling model.

1 Introduction

The aim of this study was to detect trending topics in web communications among consumers using social media, with a focus on topics related to new products. This was done by classifying words into clusters based on their co-occurrence. We collected Twitter entries about new products based on their specific expressions of sentiment or interest. Twitter is an online social networking and microblog service that enables users to post and read tweets, which are text-based messages of up to 140 characters in length. Twitter has been spreading recently in Japan. To help identify market trends, analysis of consumer tweet data has received much attention. In this study, we examined temporal variation in topics regarding new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. Twitter is an online social networking and microblog service that enables users to post and read text-based messages, known as tweets. Although a single Twitter entry is limited to 140 characters, this is sufficient to express ideas and even to write a short story in the Japanese language. In Japanese, just a few

A. Nakayama
Tokyo Metropolitan University, 1-1 Minami-Ohsawa, Hachioji-shi 192-0397, Japan
e-mail: atsuho@tmu.ac.jp

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_13

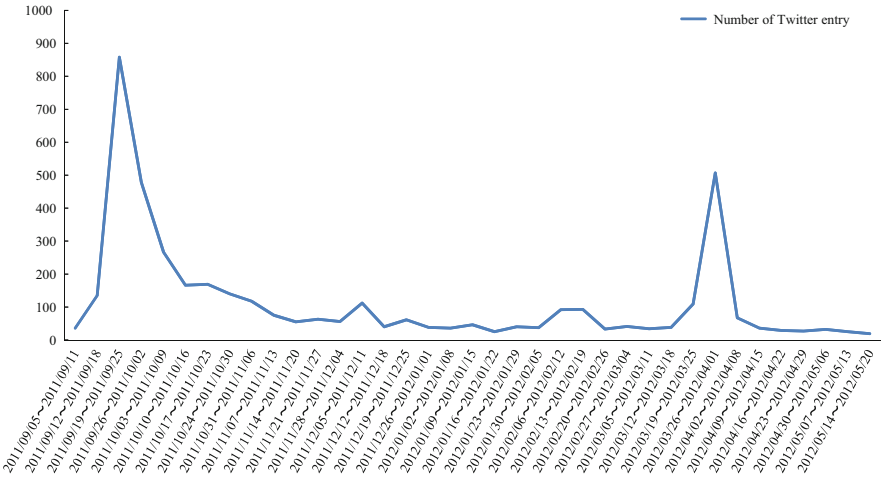

Fig. 1 The weekly number of Twitter entries regarding a new type of inexpensive, beer-like beverage

characters can convey a considerable amount of information. The Japanese writing system uses characters derived mainly from Chinese characters. For example, "経済 (keizai)," which means "economics" in Japanese, consists of two characters, whereas the equivalent English word contains nine. Thus, the limitation on the number of characters per tweet is less restrictive in Japanese, stimulating tweet posting in Japan. Tweeters can easily post short entries at any time and from any location using their smartphones or other mobile devices.

It is important to consider the temporal variation in trending topics when detecting such trending topics by classifying words into clusters based on co-occurrence of words in Twitter entries. Personal concerns are influenced by product strategies, such as marketing communication strategies, and thus change over time. For example, Fig. 1 shows the weekly number of Twitter entries regarding a new product, an inexpensive, beer-like beverage. On September 22, 2011, a new product was launched, and TV commercials for the product began running that week. A peak in Twitter entries was reached a few weeks after the product release. The number of entries per week slowly decreased after this peak. The gross rating point declined during the 2 months following the release, and the number of entries also decreased. Small peaks, however, were triggered by the release of new TV commercials. These data show how the weekly number of Twitter entries exhibited temporal change. Thus, to understand topic characteristics, it is important to consider temporal variation in trending topics and to establish criteria to select appropriate words that are representative of such temporal variation. We chose keywords representing various topics from Twitter entries and tracked the weekly variation in these topics. We then classified the words extracted from Twitter data using non-negative matrix factorization as a dimensionality reduction model. Finally, we used a visualization method for text classification to interpret the results, employing a multidimensional scaling model.

Table 1 An example entry × word matrix

	heat	hot	is	next	persists	the	today	very	week
Entry 1	0	1	1	0	0	0	1	1	0
Entry 2	1	0	0	1	1	1	0	0	1

2 The Data

The text data of Twitter entries regarding certain product names were searched and collected at 5-min intervals. We created a system for data cloning that was programmed in Ruby (https://www.ruby-lang.org/en/). Due to changes in specifications of Twitter API, our system is not currently operational. However, as of recently, we are able to easily collect Twitter entries by using the R package "twitterR," and libraries are updated and released in many programming languages. We searched for Twitter entries regarding a new brand of inexpensive, beer-like beverage named "金のオフ" (Kin no Off) produced by Sapporo Breweries, Ltd., and we collected 4622 tweets from September 2, 2011 through May 18, 2012. Our reasons for focusing on Twitter entries regarding new beverage products were twofold: it is a useful means of diminishing the effect of past product strategies such as merchandising and advertising, and we have found it particularly easy to evaluate time series variation in personal concerns for newly released beverage products.

In this study, we looked for topics associated with new products by classifying words into clusters based on the entry × word matrix of Twitter entries. For example, the entry × word matrix shown in Table 1 consists of the following nine words: heat, hot, is, next, persists, the, today, very, and week. Entry 1 is "Today is very hot," and Entry 2 is "The heat will persist next week." To detect topics more easily, we tokenized each tweet message that was written in sentences or sets of words. However, one of the most difficult natural language-processing problems in Japanese is tokenization. This is referred to as the "wakachigaki" problem. In most Western languages, words are delimited by spaces and punctuation. In Japanese, words are not separated by spaces. Consider the following sentence, "今日はとても暑い" (kyouhatotemoatsui). The English translation of this sentence is, "Today is very hot." In contrast, there are no spaces or separation symbols between the Japanese words. We used morphological analyses such as tokenization, stemming, and part-of-speech tagging to separate the words as follows.

今日	は	とても	暑い
kyou	ha	totemo	atsui
noun	Japanese particle	adverb	adjective

In our study, we used the Japanese morphological analyzer ChaSen to separate words in passages and to distinguish all nouns, verbs, and adjectives. ChaSen (http://chasen.naist.jp/) is a fast, customizable Japanese morphological analyzer that takes

Table 2 Example of a dataset to demonstrate the method used to calculate CSM

	Week *i*	Week *j*
Frequency of the word X	*a*	*b*
Frequency of words other than X	*c*	*d*

the form of a hierarchical structure. It is designed for generic use, and can be applied to a variety of language-processing tasks. A detailed discussion of ChaSen can be found in [6].

Next, we selected keywords representative of our chosen topics. To better understand topic characteristics, it was important to establish criteria to choose appropriate words representing temporal variation. We performed a statistical analysis based on the complementary similarity measure (CSM; [9]). CSM has been widely applied in the area of character recognition, and was originally developed for the recognition of degraded machine-printed characters. To construct appropriate word-set topics each week, we estimated the associations within word pairs. CSM is able to measure the inclusion relation between weeks *i* and *j* to recognize characters and identify word trends on a weekly basis. Given the following table of data, CSM is defined as follows:

$$CSM(Week\ i, Week\ j) = (ad - bc) \Big/ \sqrt{(a+d)(b+c)}. \tag{1}$$

CSM is an asymmetric measure (Table 2). Chi-square values have often been used to estimate the relation between two words, and are defined as follows:

$$\chi^2 = N(ad - bc)^2 \Big/ \sqrt{(a+b)(c+d)(a+c)(b+d)}, \tag{2}$$

where $N = a + b + c + d$. The formulas for CSM and chi-square are quite similar. However, the chi-square analysis is more likely to select words occurring with low frequency compared to the CSM method when analyzing data that contain a large spread in the occurrence frequency of words. Certain words occurred only rarely, whereas others occurred quite frequently in the text data of Twitter entries used in this study. Thus, the frequency of occurrence of some words was hundreds of times larger than that of others. For this reason, we decided to use CSM in this study. We collected the words receiving the top 10 CSM scores each week, and retained words with a total selection frequency of eight or more. The CSM score depends on word frequency. Thus, it was possible for words with low total frequency of occurrence to be selected as distinct words during a particular week, provided that the words occurred frequently that week. We extracted 359 words and removed all entries that did not include any of these words. The dataset comprised 4232 entries × 351 words. The data showed co-occurrences among 351 words in the selected entries.

3 The Analysis

The entry × word matrix obtained from the Twitter entries was sparse and of high dimensionality, so it was necessary to perform a dimensionality reduction analysis. We employed some excellent computing resources to help analyze the highly dimensional and sparse matrices. In addition, these matrices often contained noise, making it difficult to uncover the underlying semantic structure. Because of these difficulties, we found it necessary to implement dimensionality reduction. To reduce dimensionality, procedures such as Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) [2] and Probabilistic Latent Semantic Analysis (PLSA) or Probabilistic Latent Semantic Indexing (PLSI) (e.g., [5]) are often applied. LSA reduces the dimensionality of the entry × word matrix by applying a singular value decomposition (SVD), and it then expresses the result in an intuitive and comprehensible form. However, it can take a long time to perform LSA on a large matrix. In PLSA, a probabilistic framework is combined with LSA. This method uses mixture decomposition (the convex combination of aspects), which has a well-defined probability distribution. The factors have clear probabilistic interpretations in terms of the distribution of mixture components. We analyzed the entry × word matrix using Non-negative Matrix Factorization (NMF) to reduce the dimensionality [8]. Similar to principal component analysis (PCA), NMF consists of positive coefficients in linear combination. The computation of NMF is based on a simple iterative algorithm, which is particularly useful for applications involving large, sparse matrices. Ding et al. [3] have shown that both NMF and PLSI (PLSA) optimize the same objective function, ensuring that the use of NMF and PLSI is equivalent.

NMF is used for dimensionality reduction as follows:

$$V_n \approx W_n \times H_r, r < nm/(n+m) \tag{3}$$

The matrix V consists of non-negative data, such as that in an entry × word matrix. The matrix W contains non-negative basis vectors and shows the strength of associations between words and topics. The matrix H contains non-negative coefficients and shows the strength of associations between entries and topics. We can detect topics involving new products using the basis vector coefficients. The results are conceptually similar to those of PCA, but the basis vectors are non-negative. Here, the original data are represented purely through additive combinations of the basis vectors. This characteristic of NMF, i.e., data representation based on additive combinations, is effective because it suggests the intuitive notion of combining parts to form a whole. NMF computation is based on this simple iterative algorithm, and it is very efficient for applications involving large matrices.

Personal concerns are influenced by new product strategies, such as marketing communication strategies, and they change over time. It is important to consider the temporal variation in trending topics when detecting trending topics by classifying words into clusters based on co-occurrence of words. To assist us in the

interpretation of the effects of temporal variation, we visualized the results of text classification using multidimensional preference scaling (MDPREF; [1]), which provided SVD of the scalar products of the preference ratings data. MDPREF provided a "point-vector" representation, such that the columns (i.e., stimuli) were represented as points, and the rows (i.e., subjects) as unit vectors. We revealed temporal variation by analyzing the coefficients in the matrix H, regarding these associations among entries and topics as preference data.

Lattin et al. [7] have formalized the MDPREF model as follows. Let s_{ij} denote the preference expressed by an individual i for a stimulus j. According to the vector model, the subjective utility can be represented by

$$S_{ij} = y_i' x_j \tag{4}$$

where y_i' is a row vector representing the relative preference of individual i, and x_j represents the location of objects j in multidimensional space. We can write the model in matrix form as follows. According to the vector model, the subjective utility can be represented by

$$S = YX' \tag{5}$$

where Y is a matrix with m rows (one for each individual), X' is a matrix with n columns (one for each object), and S is the $m \times n$ matrix of subjective utilities. The rows of Y can be normalized to unit length so that they correspond to unit vectors. If we assume that the subjective utilities expressed by the individual have metric properties, then we can solve for Y and X' by factoring S using matrix decomposition. According to the vector model, the subjective utility can be represented by

$$S = U\beta V' \tag{6}$$

We then set $Y{=}U\beta$ and $X{=}V$, defining X and Y that yield the best ordinary least squares r-dimensional vector model representation of the S matrix. Carroll [1] points out that this procedure, outlined by Eckart and Young [4], produces matrices X and Y such that $\hat{S}{=}YX'$ is indeed the best ordinary least squares r-dimensional vector model approximation of S.

3.1 Analysis of Topic Classification

Note that throughout this section, Japanese words will be followed by their English translations in parentheses. We classified the words extracted from the tweet data regarding a new brand of inexpensive, beer-like beverage named "金のオフ" (Kin no Off) produced by Sapporo Breweries, Ltd. We implemented NMF to reduce dimensionality using an R package "NMF" based on Lee's model [8]. "Kin no Off"

contains 50% less purine and 70% less carbohydrates than other inexpensive, beer-like beverages. It is thus classified as a third-category beer, containing ingredients such as corn, soybeans, and peas rather than malt for the purpose of price reduction. For Japanese taxation purposes, brewed malt beverages in Japan fall into one of three categories: beer, Happoshu, or third-category beer. Alcoholic beverages made from malt are classified as beer if their malt content exceeds 67%. If a beverage contains less than 67% malt content, it falls under the tax category of Happoshu. Japanese breweries have produced even lower-taxed and non-malt brews made from soybeans and other ingredients, which do not fall under either of these classifications. These lower-taxed, non-malt brews, referred to by the mass media as third-category beers, were developed to compete with Happoshu.

Lee's model is an algorithm based on Euclidean distance that uses simple multiplicative updates. We determined that the maximum number of topics was 10, and the minimum as 4. In this analysis, eight topics are discussed for interpretation purposes. Table 3 lists the eight topics and shows the top 10 heavily weighted words in the basis vector W. Spellings using the Roman alphabet as well as English translations of the Japanese words are also shown in Table 3. From results such as these, we are able to identify the one or two words that are most heavily weighted. As a result of Twitter's 140-character limit, each topic consists of a small, core set of words. We can detect the prevalence of certain topics based on observations of which words are most heavily weighted.

We were able to divide the eight topics into three groups. One was the review topics, which consisted of Topics 1, 3, 7, and 8. Topic 1 was the review containing a link to an external website and product images, Topic 3 was the review of purchasing behavior and information about the new product, Topic 7 was the review of the brewery's release of the new product, and Topic 8 was the review of experiences actually drinking the product. The second group was the topics associated with advertising, which consisted of Topics 2, 5, and 6. Topic 2 was about advertisements on the train, Topic 5 was about TV commercials, and Topic 6 was concerned with performers in TV commercials. The third group consisted only of Topic 4. Topic 4 was not associated with inexpensive beer-like beverages, and the product name used as a keyword to extract Twitter entries that occurred in a different context.

Topics 1, 3, 7, and 8 are all based on reviews, though in various ways. In Topic 1, the words associated with the review containing a link to an external website and product images are heavily weighted. The most heavily weighted word was "http" (http), so it is the core word of Topic 1. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. Some Twitter entries were posted containing links to the external website and in-line product images, as well as phrases such as "the new product 'Kin no Off' was released, and I updated my blog about it"; "the new TV commercial for the new product 'Kin no Off' was broadcast"; or "'Kin no Off' is a new release and it tastes good." Some users posted links to external websites, such as their own blogs or the manufacturer's homepage. Others added in-line product images to their tweets. We therefore believe that it would be possible to infer the topic of these tweets, namely reviews containing a link to an external website as well as product

Table 3 The eight topic results and the top 10 weighted Japanese words in the basis vector W

Topic 1				*Topic 2*			
Japanese	*Roman alphabet*	*English translation*	*Weight*	*Japanese*	*Roman alphabet*	*English translation*	*Weight*
ｈｔｔｐ	ｈｔｔｐ	http	0.48	可愛い	kawaii	cute	0.28
オフ	ofu	off	0.02	広告	koukoku	advertisement	0.21
更新	koushin	update	0.02	見る	miru	see	0.08
ブログ	blog	blog	0.02	良い	yoi	good	0.05
良い	yoi	good	0.02	電車	densha	train	0.05
発売	hatubai	release	0.01	永作	Nagasaku	Nagasaku	0.04
なう	nau	now	0.01	電車内	denshanai	on the train	0.03
ひる	hiru	daytime	0.01	人	hito	people	0.03
新ＣＭ	shinCM	new TV commercial	0.01	ー	ー	ー	0.02
新発売	shinhatubai	new release	0.01	ω	ω	ω	0.02
Topic 3				*Topic 4*			
Japanese	*Roman alphabet*	*English translation*	*Weight*	*Japanese*	*Roman alphabet*	*English translation*	*Weight*
ビール	biiru	beer	0.13	ＲＴ	RT	RT	0.21
買う	kau	purchase	0.08	なう	nau	now	0.04
オフ	ofu	off	0.07	w	w	w	0.03
美味しい	oisii	delicious	0.06	金	kin	Friday	0.02
プリン体	purintai	purine	0.04	オフ会	ofukai	alcoholic party	0.02
味	azi	taste	0.04	予定	yotei	schedule	0.02
50%	50%	50%	0.04	いる	iru	stay	0.02
発泡酒	happoushu	low-malt beer	0.04	下さる	kudasaru	do	0.02
上手い	umai	tasty	0.03	お願い	onegai	please	0.02
糖質70	toushitu70	carbohydrate 70	0.02	宜しい	yoroshii	kind regards	0.02
Topic 5				*Topic 6*			
Japanese	*Roman alphabet*	*English translation*	*Weight*	*Japanese*	*Roman alphabet*	*English translation*	*Weight*
ＣＭ	CM	TV commercial	0.51	永作	Nagasaku	Nagasaku	0.56
見る	miru	see	0.06	可愛いすぎる	kawaiisugiru	way too cute	0.12
ー	ー	ー	0.03	見える	mieru	appear	0.03
可愛い	kawaii	cute	0.03	ポスター	Posuta	poster	0.02
出る	deru	perform	0.03	大島優子	Oshima Yuuko	Yuuko Oshima	0.02
似る	niru	resemble	0.03	似る	niru	resemble	0.02
やる	yaru	do	0.02	好き	suki	like	0.01
好き	suki	like	0.02	車内広告	shanaikoukoku	advertisement on the train	0.01
パフ	Pafu	Puff	0.01	男装	dansou	dressing as a man	0.01
曲	kyoku	music	0.01	ひる	hiru	daytime	0.01

(continued)

Table 3 (continued)

Topic 7				*Topic 8*			
Japanese	*Roman alphabet*	*English translation*	*Weight*	*Japanese*	*Roman alphabet*	*English translation*	*Weight*
サッポロ	Sapporo	Sapporo	0.48	飲む	nomu	drink	0.43
上手い	umai	tasty	0.02	寝る	neru	sleep	0.02
味	azi	taste	0.01	見る	miru	see	0.02
発泡酒	happoushu	low-malt beer	0.01	味	azi	taste	0.02
出る	deru	release	0.01	ー	ー	ー	0.02
金麦	Kinmugi	Kinmugi	0.01	美味しい	oishii	delicious	0.02
発売	hatubai	release	0.01	好き	suki	like	0.01
ひる	hiru	daytime	0.01	笑	wara	laugh	0.01
良い	yoi	good	0.01	なう	nau	now	0.01
こだわる	kodawaru	pursue	0.01	いる	iru	stay	0.01

images, solely from the most heavily weighted words of Topic 1. In Topic 3, the words associated with the review of purchasing behavior and information about the new product were heavily weighted. The most heavily weighted word was “ビール” (beer), followed by “買う” (buy). These words are the core words of Topic 3. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. “Kin no Off” contains 50% less purine and 70% less carbohydrates than other third-category beers, and is thought to be a healthier product. We believe that these features of the new product can be inferred from the list of heavily weighted words of Topic 3. Actual Twitter entries corresponding to this topic include “I bought the third-category beer named ‘Kin no Off,’ and it features 50% reduced purine and 70% reduced carbohydrate”; “the catch-phrase of the third-category beer named ‘Kin no Off’ is that it is delicious, though the purine and carbohydrate are reduced, so we should purchase it if its taste is as delicious as that of low-malt beer or especially normal beer”; and “the features of the third-category beer named ‘Kin no Off’ include are 50% reduced purine and 70% reduced carbohydrate, and it is as tasty as normal beer.” To repeat, Topic 3 is reflected in tweets concerning purchasing behavior and information about the new product. In Topic 7, the words associated with the review of the release of the new product from Sapporo Breweries, Ltd., are heavily weighted. The most heavily weighted word is the Brewery’s name, “サッポロ” (Sapporo), and it is the core word of Topic 7. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. “金麦” (Kinmugi) is a rival third-category beer. Further Twitter entries include “Sapporo ‘Kin no Off’ is tasty, and the taste is better than other low-malt beers, so I think the materials to make it were selected carefully”; and “I made a trial purchase of Sapporo ‘Kin no Off’ that had been newly released, and its taste is good.” To repeat, Topic 7 is associated with the release of the new product from Sapporo Breweries, Ltd., and reviews of its taste. In Topic 8, the words associated with reviews of drinking the product have the heaviest weight. The most heavily weighted word is “飲む”

(drink), and it is the core word of Topic 8. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. Some examples of Twitter entries associated with Topic 8 include "I like to drink 'Kin no Off'"; "personally, it is my very favorite taste"; "I will sleep well after drinking 'Kin no Off' because I am tired today"; and "I drank 'Kin no Off,' and it was more delicious than other third-category beers." To repeat, Topic 8 is associated with reviews of product consumption.

Topics 2, 5, and 6 are associated with advertising. Topic 2 regards advertisements on the train, Topic 5 is associated with TV commercials, and Topic 6 concerns a TV commercial performer. In Topic 2, the words associated with advertisements on the train have the heaviest weight. The most heavily weighted word is "可愛い" (cute), followed by "広告" (advertisement). These words are the core words of Topic 2. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. Hiromi Nagasaku (永作 博美), a popular Japanese actress, appeared in the advertisements on the train. We believe that it would be possible to infer this by observing the top words of Topic 2. In Topic 5, the words associated with advertising have the heaviest weight. The most heavily weighted word is "CM" (TV commercial), and it is the core word of Topic 5. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. The Twitter entries generally contained positive feedback in regard to the performer in the TV commercial. The song "Puff, the Magic Dragon" played during the TV commercial, and Twitter entries addressing the music were also posted. We believe that the top words of Topic 5 reflect tweeters' impressions of the performer and music in the commercial. In Topic 6, the words associated with the performer in the advertisement are most heavily weighted. The most heavily weighted word is "永作" (Nagasaku), the name of the performer, followed by "可愛すぎる" (way too cute). These words are the core words of Topic 6. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. In the advertisement, Nagasaku is dressed as a man. Twitter entries regarding this topic have generally been positive, and have included phrases such as "I like the TV commercial performer 'Nagasaku'"; "the TV commercial performer 'Nagasaku' dressed as a man in the advertisement is cute"; or "the TV commercial performer 'Nagasaku' resembles 'Yuuko Oshima' (a popular Japanese actress and singer)." Therefore, we believe that tweeters' general impressions of Nagasaku in the commercial can be inferred by observing the list of top words for Topic 6.

3.2 Analysis of Topic Predictions

We identified weekly variation in topics by analyzing the averaged data using MDPREF. The weight coefficients indicating entries' contributions to topics were averaged on a weekly basis. This analysis used the maximum dimensionality of categories three through seven. The largest variance accounted for (VAF) in assessing

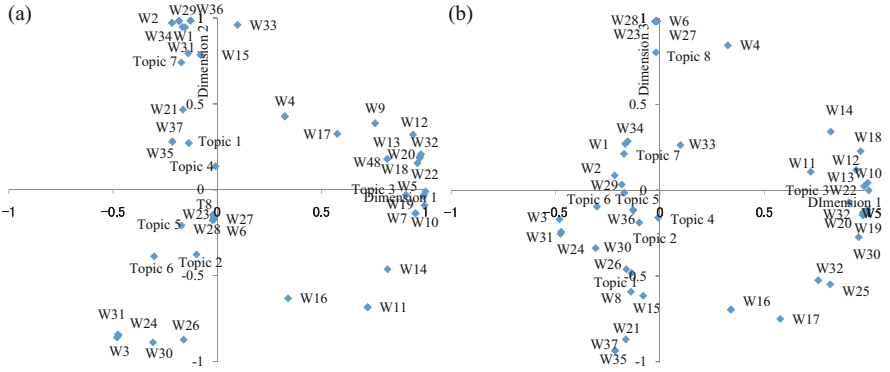

Fig. 2 Two-dimensional view of the three-dimensional configuration. The left-hand figure (**a**) defines a plane using a two-dimensional configuration for dimensions 1 and 2 derived from the results for three dimensions. The right-hand figure (**b**) defines a two-dimensional configuration for dimensions 1 and 3 derived from the results for three dimensions

the goodness of fit in each dimensionality space was selected as the maximum VAF for that space. The VAF represents the ratio of the sum of eigenvalues to the total amount of on-diagonal elements in SS'. For example, the VAF in two dimensions represents the ratio of the sum of the two largest eigenvalues to the total amount of on-diagonal elements in SS'. The VAF ratio can be represented by a configuration of variances in the preference data s_{ij}, and it measures the goodness of fit in each dimensional space. The resulting maximum VAFs from five one-dimensional spaces were 0.889, 0.766, 0.639, 0.492, and 0.329. Examination of these five VAFs initially encouraged us to adopt four- or higher-dimensional configurations as solutions. However, we did not believe that increasing the dimensionality of space beyond three dimensions would help to understand the weekly variation in topics. Three-dimensional space is sufficient to understand the weekly variation, so we chose to use three-dimensional results for the solutions. The three-dimensional configuration of the results is represented in two parts: configurations for dimensions 1 and 2 and for dimensions 1 and 3. Figure 2a shows a two-dimensional plot for dimensions 1 and 2 of the three-dimensional results. Figure 2b shows a two-dimensional plot for dimensions 1 and 3 of the three-dimensional results. In these figures, "W" represents weeks, numbered from 1 to 37, where W1 is the first week of data collection, and W37 is the last. This result provides a "point-vector" representation such that the columns (stimuli, or topics) are represented as points, and the rows (subjects, or weeks) are represented as unit vectors. The vector of topics indicates the direction of increasing preference, and the weeks are characterized by these topics.

Topic 3, which focuses on the review of purchasing behavior and information about the new product, is found in the right half of Fig. 2a. Topic 7, which is associated with a review of the brewery's release of the new product, is found in the upper half of Fig. 2a. Topics 3 and 7 are similar in terms of their characteristics, but they were the subjects of tweets during different weeks. Note the presence of

corresponding vector weeks near Topic 3, during the period 1–3 months after the product release. During this period, consumers accepted the new product, and it became more widely recognized. Accordingly, the review of purchasing behavior and information about the new product were the subject of tweets during these weeks. On the other hand, certain vector weeks experienced gross rating point increases near Topic 7 only a few weeks after the product release. Of course, topics associated with the review of the brewery's release of the new product experienced more activity during the first few weeks following the product release because of the advertisements and marketing associated with the product launch. Topics 2, 5, and 6, which were associated with advertising reviews, are shown in the lower left quadrant of Fig. 2a. During the broadcast of new TV commercials, certain vectors of weeks experienced increased gross rating points for tweets associated with topics 2, 5, and 6. These topics associated with advertising reviews are sensitive to the amount of advertising present. Topic 1, shown in the lower half of Fig. 2b, is associated with tweets containing links. Topic 8, which relates to the review of experiences of actually drinking the product, is located in the upper half of Fig. 2b. The vector weeks occurring a few weeks to several months following the product release are located near Topic 1. This shows that topics associated with sharing links in tweets were active during the weeks following the release. Presumably, the various weeks closest to Topic 8 and other topics associated with the review of experiences of actually drinking the product correspond to the times when consumers were trying the new product.

4 Conclusion

We detected trending topics related to a new product by classifying words into clusters based on the co-occurrence of words in Twitter entries. Each topic consisted of a small set of core words. The topics of Twitter entries were divided into two categories: those associated with reviews, and those associated with advertising. These topics were further classified by the characteristics of their core words. We then detected weekly trends in topics related to new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. We found that the personal concerns and tweet contents of Twitter users were influenced by new product strategies, such as marketing communication strategies, and they changed over time.

Acknowledgements We express our gratitude to the anonymous referees for their valuable reviews. This work was supported by a Grant-in-Aid for Scientific Research (C) (No. 16K00052) from the Japan Society for the Promotion of Science. We are grateful for financial support from the 45th Yoshida Hideo Memorial Foundation. We wish to thank Video Research, Ltd., for allowing us to make use of the GRP data. We are also greatly indebted to Hiroyuki Tsurumi of Yokohama National University and Jyunya Masuda of INTAGE, Inc., for their great support and advice in analyzing data.

References

1. Carroll, J.D.: Individual differences and multidimensional scaling. In: Shepard, R.N., et al. (eds.) Multidimensional Scaling, Vol. I Theory, pp. 105–155. Seminar Press, New York (1972)
2. Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
3. Ding, C., Li, T., Peng, W.: Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence, chi-square statistic, and a hybrid method. In: Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference (AAAI'06), pp. 342–347 (2006)
4. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**, 211–218 (1936)
5. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, pp. 289–296 (1999)
6. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp. 230–237 (2004)
7. Lattin, J.M., Carroll, J.D., Green, P.E., Green, P.E.: Analyzing Multivariate Data. Thomson Brooks/Cole, Pacific Grove, CA (2003)
8. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 556–562. MIT Press, Cambridge (2000)
9. Sawaki, M., Hagita, N.: Recognition of degraded machine-printed characters using a complementary similarity measure and error-correction learning. IEICE Trans. Inf. Syst. **E79-D**(5), 491–497 (1996)