Predicting Latent Trends of Labels in the Social Media Using Infectious Capacity

Yuki Sonoda and Ikeda Daisuke

Abstract—**This paper is devoted to predicting trends on the social media. Typical methods in the literature are based on temporal changes in usage of words or phrases on the media, and try to find a rapid increase, called a burst, of them. Therefore, these methods can be applied *only after* a burst is emerging. In this paper, we propose an index, called the *infectious capacity*, to detect potential trends on the social media before they would emerge. To achieve this, we focus on *labels* and *items*, and predict trends of a label, instead of those of a target object, such as contents of a social media, where an item is a concept represented by an object and a label categories items. On a photo sharing service, for example, a photo is an object, a tag is a label, and concepts represented by a photo are items for the photo. Using labels and items, the infectious capacity for a label is defined as the ratio of the variety of items with the label to the number of occurrences of the label in given data. That is, the larger value an infectious capacity of a label is, more infectious the label is. Our experiments on real data showed that the infectious capacities for most labels are substantially constant over time. This result means that we can forecast the variety per usage for a label just after the label is used. Moreover, we found that infectious capacities for popular labels have similar values. Combined with the first result, we are able to predict latent trends before labels become popular. In fact, this is also supported by experiments on tweets, where we were able to find potentially popular hashtags, regarding hashtags as labels, before they become popular. As far as the authors know, this is the first result of future trend prediction on the social media.**

Index Terms—**Category, constant index, future trend detection.**

I. Introduction

Predicting trends has been extensively studied in many fields because we can make appropriate decisions based on the predicted trends if the prediction is accurate, although predicting trends is extremely difficult in advance [1]. Such studies include forecasting physical phenomena [2], predicting financial indexes or indicators [3]-[5], estimating the number of patients of a disease [6], and forecasting sales of cultural products, such as movies [7] and songs [8].

The rapid increase of the data on the social media, such as tweets on Twitter (https://twitter.com/), leads to a rise in the importance of studying predicting trends from data on the media. However, unlike the time series data, which has serial correlation in time series models, such as ARMA and ARIMA [9], we can not assume such models for data or trends on the social media. Therefore, typical methods to predict trends for data on the social media try to grasp trends by analyzing occurrences of words or phrases in given data [10]-[16]. For example, [11] uses mentions to movies in tweets, [16], [17] mentions about earthquakes, and [3], [6], [7] query words used at search engines.

However, such methods can predict trends only after they are emerging because such a method counts occurrences of target objects such as tweets mentioning earthquakes and outputs some of them as bursty objects if a temporal change of their occurrences is beyond a predefined threshold. Therefore, it is impossible for them to predict *future* trends.

In this paper, we propose an index, called the *infectious capacity* to detect potential trends on the social media before they would emerge. It is defined for a *label* based on *items* to which the label is attached, where an item is a concept involved in an object of the social media and a label represents a concept involved in items. On a photo sharing service, for example, a label is implemented as a tag, an object is a photo, and items for a photo represent concepts of it. If a photo includes a character of a Ninja animation then "ninja" could be an item for the photo, and "animation" and "Cool Japan" labels for the items (see Fig. 1).

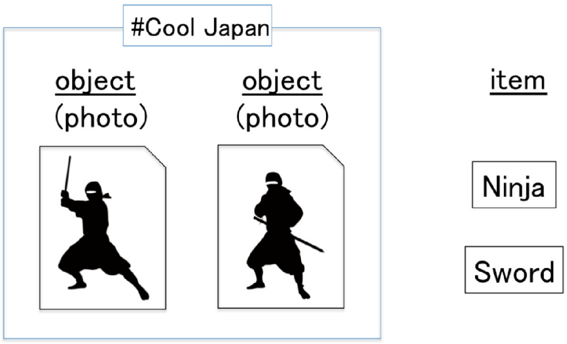


Fig. 1. On a photo sharing service, for example, a label is implemented as a tag, an object is a photo, and items for a photo represent concepts of it. If a photo includes a character of a Ninja animation then "ninja" could be an item for the photo, and "Cool Japan" labels for the items.

Using the infectious capacity, we propose a method to detect labels which will be popular. As described above, focusing on characteristics of occurrences of *target objects* for prediction do not enable to predict future trends of themselves. However, we will show focusing on characteristics of occurrences of *labels* to predict future trends of *labels*. Therefore, our focus to predict is not objects, but items.

So predicting trends of labels is also important. A label is used to classify target objects and so we can easily grasp the meaning of them via labels. Besides, a label has a wider

Manuscript received September 25, 2015; revised November 5, 2015. This work was supported by JSPS KAKENHI, Grant-in-Aid for Scientific Research (B), Grant Number 15H02787.

The authors are with the Department of Informatics, Kyushu University, Japan (e-mail: yuki.sonoda@inf.kyushu-u.ac.jp, daisuke@inf.kyushu-u.ac.jp).

doi: 10.18178/ijfcc.2015.4.6.419

influence than individual items.

This goal is quite challenging due to the following reasons: first popular labels are rare by definition; second we can not assume that occurrences of target items follow deterministic or even stochastic models. In case that we focus on a specific type of target items, such as influenza [6], songs [8], or movies [11], we can assume some knowledge about a phenomenon of trends. For example, a patient of some disease may try to search information about the disease at a search engine. However, a label is widely used for many kinds of items. Instead of assuming knowledge about specific target objects or items, we generally assume that a potentially popular label would be used for many kinds of items. In other words, the variety of items attached to a label shows potential popularity of the label.

Our contribution is three fold: firstly, we experimentally show that the infectious capacity for labels is substantially constant over time, where we use hashtags of Twitter as labels. Obtaining a constant of a system is quite interesting since, in general, such a constant shows us some important findings of the system. In addition, such a constant is useful because we can forecast the variety of a label per usage just after the label is used; secondly, we find that values of the infectious capacity for popular labels are concentrated in some range; and finally, we can predict some latent trends of labels, which means potentially popular hashtags, using thresholds defined by the found range.

II. Related Work

The researches about future or current trends prediction are divided into three types: The first type is based on a physical model. A typical method of this type is modeled by a differential equation, such as Newton's second law, which precisely shows the position of a body in the future. In other words, a method based on this type can deterministically describe current and future of target objects; the second a stochastic model. A typical model of this type is used for time series data [8]; the last type does not assume models.

For the first two types, models generally describe the evolution of some system over time. Therefore, we can predict future trends of the system using such a model. On the other hand, for the data on the general social media, we do not have an appropriate model which describes future behavior of the data. Needless to say, we do not know existence of a deterministic mechanism for the data on the social media, and we can not assume a stochastic model for the general data on the social media since the data seems not to have serial correlation. For example, given two successive tweets, there is not a correlation in general between them. In other words, a tweet does not influence the next tweet because various topics are mixed in timeline of Twitter. Without deterministic or stochastic models, it is quite difficult to forecast future trends on the social media.

Basically, methods for the social media are to predict current trends, instead of forecasting future trends. To do so, there are two types of methods. The first one is based on temporal changes in the occurrence of target objects [12], [13]. The other type assumes some correlations of the data on the social media and other data sources, and use statistical models, such a linear model, among them. For example, after an earthquake occurs, many people tweets about it [16], [17], when people are infected influenza, many people search about influenza [6], and popular movies are likely to be tweeted more frequently [11]. Basically, the later type just uses data on the social media, but predict trends for other objects, such as movies or patient of influenza. Consequently, the later type does not predict future in social media but predict future in the real world.

III. Infectious Capacity

In this section, we define an *infectious capacity* for a label after we explain the idea of it along with items and labels for the data on the social media.

To explain the idea of an infectious capacity, we give an example of a popular label "Cool Japan" at Instagram (http://instagram.com/). This label is attached to photographs which include the concepts representing Japanese culture, such as *Fujiyama*, *Ninja* and *Ukiyo-e* (see Fig. 2). Therefore, we assume that the label is virtually attached to the concepts included in photographs, though users attach a label to photographs themselves. In summary, regarding photographs as objects, an item is a concept involved in the objects and a label represents a generic concept involved in the items.

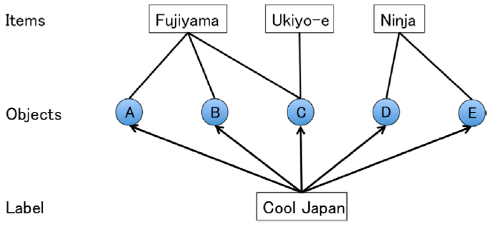


Fig. 2. Relationships with a label, objects and items on the social media.

For the label "Cool Japan", some objects are labeled with it, and "Fujiyama", "Ukiyo-e" and "Ninja" are associated with them. We can think that the label is also attached to the items.

We think that a label which has relations to many items would become popular in the future, and we introduce the infectious capacity of a label as the ratio of the variety of items with the label to the number of occurrences of the label.

Let I be a set of *items*, L a set of *labels* for items. An object o can be attached with a label l. In this case, we say that o is *labeled with* l. The object o can represent many concepts, that is items. In this case, we also say that each item is labeled with l. For a label l, we say that an item i is *labeled with* l, and an object o is associated with the item, therefore o represents the item and includes it. For an item i, we use $O(i)$ to denote the set of objects associated with i. In general, an object o can be associated with many items. If an object o is labeled with a label l, we say that an item i associated with o is also *labeled with* l.

For a label l, we also use $I(l)$ and $O(l)$ to denote the set of items and objects labeled with l, respectively. For a label l, the *frequency* $f(l)$ of the label is defined to be the number of objects labeled with l, that is, $f(l)=|O(l)|$. Similarly, for a label l, the *variety* $v(l)$ of the label is defined to be the number of items labeled with l, that is, . For a label l, the *infectious capacity* $\alpha(l)$ is defined as follows:

$$\alpha(l) = \frac{|I(l)|}{|O(l)|}. \qquad (1)$$

For example, for a label l, given $I(l) = \{i_1, i_2, i_3\}$ and $O(l) = \{o_1, o_2, o_3, o_4 \in O(i) \mid o_1 \in O(i_1), o_2 \in O(i_2) \cap O(i_3),$, $o_3 \in O(i_3), o_4 \in O(i_1) \cap O(i_3)\}$ then $|O(l)| = 4$ and $\alpha(l) = 3/4$.

Our final goal is to predict future trends of a label. Therefore, we need the infectious capacity at some time point t, denoted by $\alpha_t(l)$, for a label l, where $\alpha_t(l)$ is defined using sets of items and objects at t or before t.

If a value of $\alpha_t(l)$ for a label l is higher, the label is used for many different items, and if a value of $\alpha_t(l)$ for a label l is lower, the label is used only particular items.

IV. Experiments

In this section, we conduct experiments utilizing tweets and hashtags of Twitter. First, we explain our data set. Second, we reveal two properties of the infectious capacities of hashtags (Section B.). In Section C., thresholds about the infectious capacity for prediction are determined utilizing training data in which manually detected popular hashtags are given. In Section D., we forecast potentially popular hashtags using determined thresholds.

A. Data Set

In this section, we explain our data set and the following notes on the data:

- How to define items in tweets.
- How to process prevailed labels among collected data.

We collected tweets posted in Japan from July 6, 2012 to July 14, 2012. There exist about 146,000 hashtags and 3.5 million tweets labeled with the hashtags in the period.

We consider the infectious capacity of hashtags in tweets, regarding hashtags as labels, that is, L is a set of hashtags. In this case, a tweet is labeled with a hashtag, so O is a set of tweets and I is considered to be a set of concepts represented tweets. However, in general, it is quite difficult to automatically determine concepts included in a tweet. In our experiments, we regard common nouns in tweets as the concepts, that is, I is the set of common nouns in tweets. For example, given an object "My favorite foods are sushi and tempura #FavoriteFood", a label is *FavoriteFood*, and items are *food*s, *sushi* and *tempura*.

As preprocessing, we utilize MeCab (http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html/) to extract nouns in a tweet. MeCab is a Japanese morphological analyzer. When people make tweets, they frequently use slangs in tweets, and it is difficult for MeCab with standard dictionaries to extract them. To extract common nouns and attempt to improve accuracy of morphological analysis, we add title lists of Wikipedia and keywords of Hatena Diary (http://d.hatena.ne.jp/) as common nouns to the dictionary of MeCab. Further, we separated the Japanese words in texts with space and removed alphanumeric characters.

In a tweet, a tweet can contain two or more different hashtags. In this case, we think that the tweet is labeled with all hashtags, that is, the tweet is treated as the corresponding object for each hashtag. For example, given a tweet "My favorite foods are sushi and tempura #FavoriteFood #Cool Japan", the tweet is an object labeled with both *FavoriteFood* and *Cool Japan*.

General tags such #tbt and #nofilter can be used for many items. Too general tags have a problem in predicting future trends of labels. Considering the life cycle of such a general hashtag, we can say that it prevails after it is used for many items, and we find it as a background tag. In this case, the infectious capacity of such a background tag must be too large, so the infectious capacity can not distinguish popular tags from background tags. So we will remove them among collected data.

To remove prevailed tags, we use only hashtags which have not been used in some period and regard them as unpopular tags. We use unpopular hashtags in the training phase (resp., test phase), which were not used in the period of 12 hours before the training phase (resp., the period of the training phase) (see Fig. 3). In other words, we assume that prevailed tags appear frequently and thus must appear in this period.

As a result, in the training phase, we use tweets for about three days from July 6, and there exist about 100,000 hashtags, which were unpopular, and 350,000 tweets labeled with these hashtags. In the test phase, we use tweets of about five days and there exist about 140,000 hashtags and 400,000 tweets labeled with these hashtags (see Table I).

We need positive hashtags which will become popular in future among currently unpopular hashtags, so we regard a hashtag satisfying $|O(l)| \geq 1,000$ as popular, that is a positive hashtag in this paper.

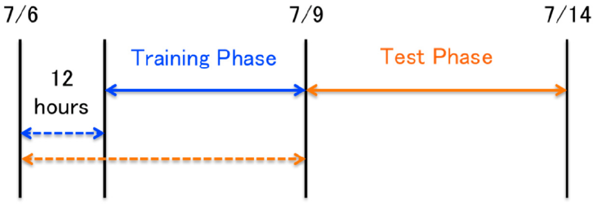


Fig. 3. This diagram shows periods we used as the training phase and test phase. We used hashtags in the training phase (resp. test phase), which were not used in the blue dotted line (resp. orange dotted line).

TABLE I: Data Set We Use in the Training and Test Phase

Phase	A period	# of tags	# of tweets
Training	7/6-7/9 (2012)	109,185	352,790
Test	7/9-7/14 (2012)	141,741	408,425

B. Properties of Infectious Capacity

We show two hypothetical properties of an infectious capacity, and then verify them. The graph in Fig. 4 shows increases of items labeled for each popular hashtag in the training phase as the number of objects used with the corresponding hashtag is increased, where X-axis $|O_t(l)|$

shows for each label and Y-axis $|I_t(l)|$. So the slope of each line in the graph represents the infectious capacity of the corresponding label. Note that, some lines have slopes whose values are almost zero.

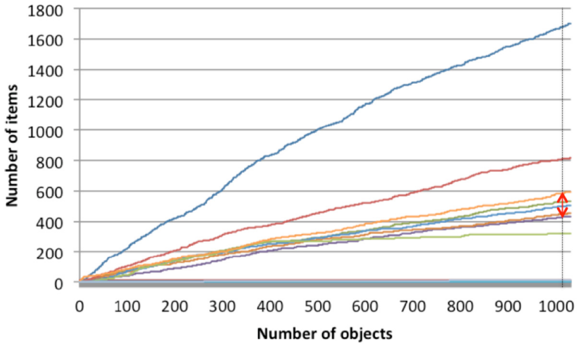


Fig. 4. This graph shows changes of the number of occurrences of items for each trendy hashtag as the number of occurrences of objects in the training phase is increase.

One line is for one hashtag. These lines look straight lines, so values of the infectious capacities are constant over time. In addition, the values of the infectious capacities of many popular hashtags are similar. The line with the highest slope is the hashtag which means "The celebrity who you have seen directly" in English. This tag is attached to many celebrities, so there exist generally many items. Therefore, the infectious capacity of the label is too high.

We find that the following two hypothetical properties from this graph.

- The slopes of the lines are almost constant, that is, the values of infectious capacities for many trendy hashtags do not change over time.
- The values of infectious capacities for many trendy hashtags are similar.

In the second property, the red arrow on the vertical line on the righthand side of the graph indicates the range of the similar values. The first property means that we can forecast the variety per usage for a label just after the label is used. Combined with the second one, we are able to predict latent trends before labels become popular using a threshold about the infectious capacity. From these properties, we can expect to predict trendy hashtags before trends and bursts emerge.

In the rest of this section, we verify the first property by single regression analysis for items and objects. Subsequently, we assume that the second property is valid, and use it for prediction in the following sections, which means we will verify the second one by the prediction result. In the analysis, explanatory variables are $|O_t(l)|$ for a label l at each time and explained variables are $|I_t(l)|$. Data for this verification is 128 hashtags in the training phase, where we use hastags satisfying $|O_t(l)| \geq 200$ in stead of $|O_t(l)| \geq 1,000$ because of statistical stability.

The mean value of coefficient of determination, which denotes the precision of the single regression analysis in the results, is 0.878. In this analysis, the p-values of all 128 hashtags are lower than 0.05, and this generally shows that the result is extremely precise. In this analysis, for a hashtag, we use the number of occurrences of it in the dataset and the number of items used with it as independent and dependent variables, respectively. Therefore, the direct consequence of the analysis is that a hashtag has a constant infectious capacity as the number of occurrences is increasing. We can extend the consequence as follows: For time points t_1 and t_2 $(t_1 > t_2)$, the infectious capacity for a hashtag is not changed whether or not it is used during this period due to the above result. So the infectious capacity is also constant over time. That is, if a hashtag has a constant infectious capacity over the number of occurrences of it, then so it does over time. Thus it result proves the first property.

Although we have shown that the infectious capacity is constant as time passes, there must be some upper bound for the number of the items which could be labeled by one label. Therefore, the above result does not mean that the value of the infectious capacity for a label is constant forever.

C. Training Phase

From the second property in the previous section, we can predict trends by using a threshold derived from the property because we can easily test if the infectious capacity of a given unknown hashtag is in the threshold. So in this section, we determine thresholds, assuming that we have a set of positive hashtags.

The second property comes from the fact that all popular hashtags have similar values of their infectious capacities. This fact is found in Fig. 4, where we have defined 1,000 to be the threshold of popularity. However, to predict future trends of labels, we need to evaluate infectious capacities at an earlier point t. Therefore, we use time points t satisfying $|O_t(l)|$ = 200, 300, 400 or 500.

In general, an arbitrary range of infectious capacities does contain negative hashtags and does not contain some positive ones. So, we evaluate each range for all pairs of hashtags by F-measure and choose the range for the threshold which achieves the highest F-measure. F-measure is a weighted harmonic average of precision and recall, and defined by $F = 1/((s/P)+(1-s)/R)$, where precision $P = |A \cap B|/|A|$, recall $R = |A \cap B|/|B|$, A is a set of hashtags predicted to be popular in the future, and B a set of positive hashtags. Although the weight s of F-measure is 1/2 in many cases, we set s= 1/4 because we put a larger priority on recall R. Trends of labels are rare phenomenons, so we place more importance on detecting trends than on making wrong prediction.

We found that there exists the hashtags which are attached many times to some specific item among positive hashtags. In other words, an item is shared among so many objects to which a hashtag are attached. Therefore, the infectious capacity of such a label is low. In the process of threshold determination, we remove the hashtags having the extremely low infectious capacity from positive hashtags in the training phase because our idea can predict the labels used for many different items.

The result is shown in Table II. Basically, the range becomes wider as $|O_t(l)|$ increases since the number of negative hashtags decrease as $|O_t(l)|$ increases.

TABLE II: EACH THRESHOLD DETERMINED AT THE NUMBER OF OBJECTS IN THE TRAINING PHASE

$\lvert O_t(l)\rvert$	Threshold
200	0.625 ~ 0.760
300	0.630 ~ 0.693
400	0.520 ~ 2.080
500	0.488 ~ 1.998

D. Test Phase

We show the result of the test phase, using the thresholds decided in the previous section. First, we show that popular hashtags can be detected by using labeled data and recall/precision criteria. Second, we show that some of the detected hashtags are detected before they become popular, that is, our method can predict trendy hashtags before they emerge.

The result is shown in Table III. In Table III, the first column represents the threshold T of popularity, the second column the number of hashtags whose frequencies are more than or equal to the number in the first column. For example, we have 44 hashtags whose frequencies are at least 500. The other columns are for evaluated values, precision, recall and F-measure, where there exist 18 positive hashtags in the test phase. Though there exist about 140,000 hashtags in total in the test phase, popular tags are very few since they are rare.

This table shows that the result basically becomes better as the number of objects is increased except that $\lvert O_t(l)\rvert = 300$. When , the threshold is narrow (see Table II), therefore the result is not good. However, the result at $\lvert O_t(l)\rvert = 300$ is much worse than at $\lvert O_t(l)\rvert = 200$ because the range of the threshold is small. At $\lvert O_t(l)\rvert = 300$, increasing the precision by setting the range of the threshold small more improve F-measure than increasing the recall by setting the range of the threshold large, as a reslut, the range of the threshold is small.

TABLE III: RESULTS OF EXPERIMENTS IN THE TEST PHASE

$\lvert O_t(l)\rvert$	# of tags at t	P	R	F
200	143	0.263	0.278	0.274
300	83	0.333	0.111	0.133
400	56	0.458	0.611	0.560
500	44	0.524	0.611	0.587

TABLE IV: TRENDY BUT NOT PREDICTED HASHTAGS, TRANSLATED FROM JAPANESE HASHTAGS

Hashtags
Retweet if you are an idiot
You show the world your shortcomings when you make a retweet
Retweet if you want the man in the Friday road show to come back
The Kanagawa Prefecture Police
No parking
Threewords
Happybirthdayheechul

Even allowing the fact that forecasting future trends without models seems to be quite difficult in general, we can not conclude that the results in Table III are good. Examining hashtags which are not predicted as trendy ones, we find that these tags are used in retweets, since all retweets of one original tweet are the same, the variety of items used with such a label is not increased. Table IV shows the hashtags which are popular but not predicted by our method. The hashtags shownd in this paper are translated from Japanese hashtags. 4 hashtags out of positive 18 hashtags are used for retweets. Therefore, the values of infectious capacities for these labels are far below under the defined threshold.

We can detect retweets since they are the same object. If more than 3/4 of tweets labeled by a hashtag are retweets, we remove the tweets attached with the hashtag. Then we evaluate again using the data set without these tags. Table V shows the result. Now, our method achieves approximately 0.7 of F-measure.

In total, 11 hashtags out of 14 positive hashtags which are not used as retweets are able to be predicted by our method. So we can conclude that our method is able to predict trends of labels, which do not have extremely low infectious capacities, with a high precision. Table VI shows predicted hashtags. Although the sentence of each hashtag translated from Japanese into English is long in Table VI, original hashtags in Japanese are short since Japanese has ideographic characters.

Next, we examine whether we can predict trends of labels, *i.e.* hashtags, before they emerge.

TABLE V: RESULTS OF EXPERIMENTS IN THE TEST PHASE AFTER 4 HASHTAGS WHICH ARE MAINLY USED FOR RETWEETS ARE REMOVED FROM POPULAR HASHTAGS

$\lvert O_t(l)\rvert$	P	R	F
200	0.263	0.357	0.328
300	0.333	0.143	0.167
400	0.458	0.786	0.667
500	0.524	0.786	0.698

TABLE VI: TRENDY HASHTAGS TRANSLATED FROM JAPANESE HASHTAGS WE WERE ABLE TO PREDICT IN THE TEST PHASE

Label	Hashtags
1	When you attach “of love” to Japanese history terms, you become popular
2	ozawa-shintou (name of a political party)
3	You may be able to understand your true character by looking at the predictive text of your cell phone
4	When you type “an arrest”, aim for the predicate text to type “a release”
5	If you are talking about me, what am I so famous?
6	If I were a teacher, I would give students homework like this over summer vacation
7	I am not going to get mad, so please tell me what you thought of my first impression
8	If you type “me”, you will receive the first set of predictive text
9	If you simplify your name to just your initials, your name become like “delinquent”
10	RPG Gamer Test
11	My follower introduced me to three new words

The graph of Fig. 5 shows increases of the number of objects for the predicted labels over time, where Y-axis shows the total tweets, that is $\lvert O_t(l)\rvert$, X-axis time t, and one unit of time is 5 minutes. Two graphs of Fig. 6 are obtained by magnification near the origin of Fig. 5, where both graphs are the same but, on the lower graph, the labels in red (resp., blue) are predicted as popular at $|O_t(l)| = 200$ (resp., $|O_t(l)| = 400$). We can see that each hashtag is predicted before labels emerge or are emerging. For example, label 7 has been predicted at before the trend of it emerged at. So we can conclude that we have predicted latent trends of labels using an infectious capacity. These graphs show the comparison with conventional techniques detecting trends after they emerge.

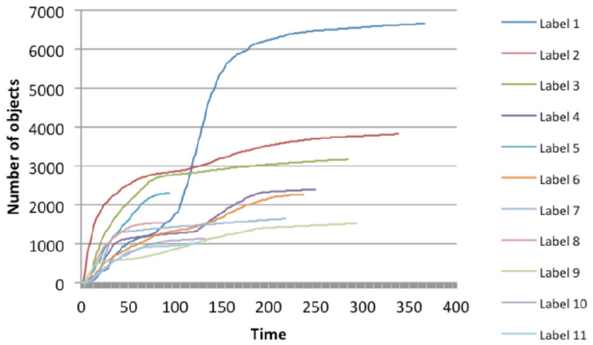


Fig. 5. This graph shows changes of the number of objects for popular labels over time.

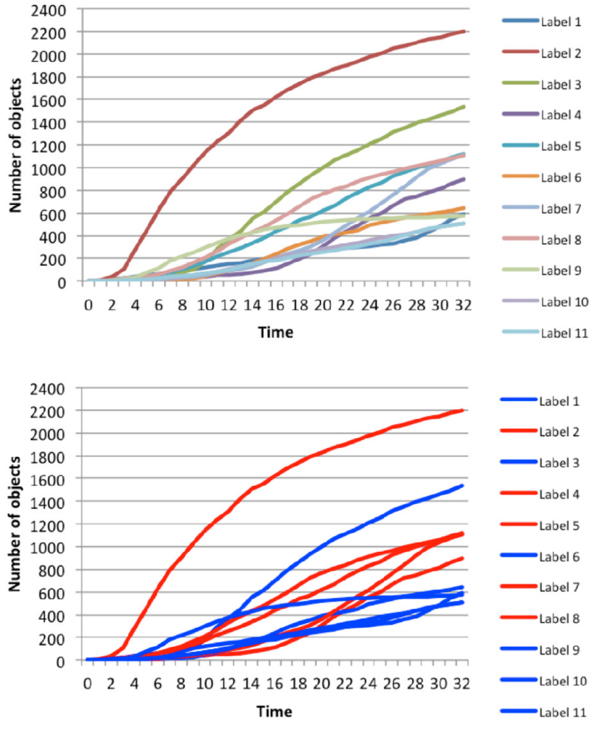


Fig. 6. Both graphs are obtained by magnification near the origin of Fig. 5, where all popular hashtags are shown in the two graphs, but on the lower graph, the labels in red (resp. blue) are predicted as popular at $|O_t(l)| = 200$ (resp. $|O_t(l)| = 400$).

V. Conclusion

A big goal of our research is to develop a method to predict future trends before trends are emerging. To do so, we have considered the trend of labels, instead of items, treating hashtags as labels. Focusing on relations with labels, items and objects, we have introduced the infection capacity of labels, and proposed a method to predict trends without building a model and without using temporal changes of the occurrences of labels. Using this method, we have predicted latent trends of labels before labels become popular.

Surprisingly enough, we have experimentally discovered that the infection capacity of labels do not change at any time and maintain constant values. This results seems to be contrary to intuition.

We have experimentally proved that our method can predict trendy hashtags before they emerge. We did not compare our method directly to other existing method because of the following reasons. First of all, existing methods, such as burst detection, can be applied after hastags become popular. Secondly, it is arbitrary to choose an appropriate value for a parameter of existing methods which use the parameter for gradient of the number of occurrences of hastags over time to detect a rapid increase of hashtags. However, the numbers of occurrences of both popular and unpopular hashtags go up and down dynamically, and increasing or decreasing rate depends on hashtags' features. For example, even when hashtags for national events have larger occurrences than those for local events, the gradient for hashtags for local events can be much larger than that for national events. Thus, to set a threshold for the gradient, we must know what kinds of events used for hashtags. On the other hand, our method is not influenced by gradient of the number of occurrences of hashtags and does not require knowledge about events (items).

Although a popular event or something important will affect the popularity of a label, our method concern the use of a label and predicting whether a label is used. So we do not need to know what an event for a label is. For example, given the hashtag #Brazil, though events such as the Olympics in Brazil will affect the popularity of #Brazil, we are not predicting popular events. Regardless of the kinds of events, all we have to do is counting the number of events labeled with #Brazil.

Besides, our method does not only predict a label which will be trendy shortly after it is created, but we does predict one which will be used many times in the future.

Since the infectious capacity is simply defined with the variety and frequency of items, we can expect that the proposed index is applicable to a wide variety of targets in many fields. However, preliminary experiments using hashtags and tweets in English do not show that the index can be used to predict trendy hashtags in English. We think that this is because of the following reasons. First, Japanese has ideographic characters and so we can express out intention in the small number of characters. So they use complex hashtags for tweets with the limited number of characters. Second, some usage of hashtags in Japan is different from that in other cultures. We sometimes use hashtags for a word game, called *Ogiri*, where, given a topic, players try to make an answer with wit and humor. For example, given "CalmDownABand", an answer might be "Water Guns'N Roses". So for making intriguing answers, various concepts are involved, and therefore, many items are used for answers. However, there are not many such a situation in English. Therefore, it is an important future work to apply the proposed index to data in different language.

Although we can predict popular hashtags which are used for many items, the labels peculiar to some specific items can not be detected. In other words, there are another type of popular labels. From preliminary experiments, the values of infectious capacities for such labels also have some range. So, detecting such a label is also an important future work.

It is also an important future work to evaluate our method using a large amount of data and improve precision of trends detection.

References

[1] D. Sornette and D. Zajdenweber, "Economic returns of research: The pareto law and its implications," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 8, no. 4, pp. 653–664, 1999.

[2] T. Tokunaga, D. Ikeda, K. Nakamura, T. Higuchi, A. Yoshikawa, T. Uozumi, A. Fujimoto, A. Morioka, K. Yumoto, and CPMN group, "Onset time determination of precursory events of singular spectrum transformation," *International Journal of Circuits, Systems and Signal Processing*, vol. 5, pp. 46–60, 2011.

[3] H. Choi and H. R. Varian, "Predicting the present with google trends," *The Economic Record*, vol. 88, no. s1, pp. 2–9, 2012.

[4] A. Harvey, *The Econometric Analysis of Time Series*, MIT Press, 1990.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[6] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009.

[7] S. Goel, J .M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "What can search predict?" Working Paper, 2009.

[8] Y. Ni, R. Santos-Rodriguez, M. McVicar, and T. D. Bie, "Hit song science once again a science?" presented at 4th International Workshop on Machine Learning and Music, 2011.

[9] A. Harvey, *Time Series Models*, Philip Allan, 1981.

[10] T. Althoff, D. Borth, J. Hees, and A. Dengel, "Analysis and forecasting of trending topics in online media streams," *CoRR*, 2014.

[11] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 492–499, 2010.

[12] J. Kleinberg. "Bursty and hierarchical structure in streams," in *Proc. the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 91–101, 2002.

[13] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the twitter stream," in *Proc. the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 1155–1158, 2010.

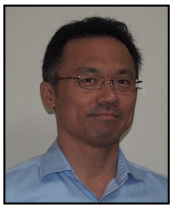
[14] S. Nakajima, J. Zhang, Y. Inagaki, and R. Nakamoto, "Early detection of buzzwords based on large-scale time-series analysis of blog entries," in *Proc. the 23rd ACM Conference on Hypertext and Social Media*, pp. 275–284, 2012.

[15] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Event extraction using behaviors of sentiment signals and burst structure in social media," *Knowledge and Information Systems*, vol. 37, no. 2, pp. 279–304, 2013.

[16] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proc. the 19th International Conference on World Wide Web*, pp. 851–860, 2010.

[17] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan, "OMG earthquake! Can twitter improve earthquake response?" *Seismological Research Letters*, vol. 81, pp. 246–251, 2010.

Yuki Sonoda was born in Japan in 1990. He received his bachelor degree in science from Kyushu University in 2014. He is currently a master course student in the Department of Informatics, Kyushu University, Fukuoka, Japan. His research interests include trend prediction and economy.



Daisuke Ikeda was born in Japan in 1971. He received his bachelor degree, master degree, and doctor degree in science from Kyushu University in 1994, 1996, and 2004, respectively.

He is currently an associate professor in the Department of Informatics, Kyushu University, Fukuoka, Japan. Formerly, he worked at Computer Center, Kyushu University, and Kyushu University Library. The latest publications include articles in the field of data mining, e-science, bioinformatics. His research interests include data analysis, such as data mining and machine learning, and data infrastructure, such as database and information retrieval.

Dr. Ikeda is serving and has served as PC members at conferences in data mining, such as International Conference on Advanced Data Mining and Applications (ADMA), Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and Discovery Science (DS). He is a member of Association for Computing Machinery and Information Processing Society of Japan.