Contents Popularity Prediction by Vector Representation Learned from User Action History

Naoki Nonaka, Kotaro Nakayama, Yutaka Matsuo
Graduate School of Engineering
University of Tokyo
Tokyo, Bunkyo 7-3-1
Email: {nonaka,k-nakayama,matsuo}@weblab.t.u-tokyo.ac.jp

***Abstract*—The anime and manga industry is important in Japan, and its popularity has been increasing overseas in recent years. Under such circumstances, predicting the popularity of media contents is important for content holding companies. Popularity prediction research has, so far, rarely considered the multifaceted information of media contents based on consumer preferences. In this study, we extracted users' preferences from Wikipedia and obtained a vector representation with multifaceted content information. We qualitatively analyzed learned vector representations and showed that accuracy is improved by 2 to 3 % in a popularity prediction task.**

Keywords–Popularity prediction, Wikipedia, Vector representation, MLP

I. Introduction

The anime and manga industry is important in Japan, and its popularity has been increasing overseas in recent years. In 2016, the market size of the animation industry grew at a large rate of 12.0% over the previous year [1]. In addition, it is expected that the scale of overseas content markets will continue to expand [2].

Under such circumstances, popularity prediction of media content is an important task for companies considering secondary use of content and content holders. If overseas copyright buyers can obtain accurate popularity information regarding media contents and information useful for predicting trends, promotion of content work to overseas will be enhanced. Predicting the number of product units sold and the future popularity of a product is important for company decisions such as marketing [3], and many studies have been conducted in this regard [4]-[5]. Representative research includes research predicting movie sales [4][6] and predicting future stock prices [5]. In research on popularity prediction of media contents, Hozumi [7][8] made a prediction using search query volume in search engine, Twitter and Wikipedia data. These studies use information regarding social media that has developed rapidly in recent years, for the purpose of prediction. In addition, accuracy is improved by using consumers' word-of-mouth information for predicting subjects in future prediction [9].

When predicting product sales, multifaceted information considering multiple degrees of similarity, such as product categories, based on consumer preferences, is equivalent in importance to consumers' word-of-mouth information. Collaborative filtering, a typical method in product recommendation, is a model that reflects user preferences, such as recommending based on a product and user vector representation [10], [11]. Thus, in the recommendation problem, a model that considers consumer preference has achieved admirable results. In the popularity prediction task, popularity prediction of movies using features obtained from word-of-mouth information [4] and popularity prediction using feature quantities extracted from prediction targets [12] existed, but there is little research in which the degree of similarity was considered for multiple scales based on a user's preference toward objects.

In this study, we aim to extract users' preferences from Wikipedia and obtain multifaceted information, such as genre and fashion age, regarding media contents. Wikipedia, which covers a wide range of contents, is one of the social media that are used and edited by many users. In Wikipedia, since the user edits pages relating to items of high interest to them, the items edited by each user are considered to reflect user preferences. Therefore, by considering the preferences of various users obtained from their editing histories, multifaceted information can be obtained for content works, taking similarity into account for multiple scales such as genre of contents or fashion age.

In summary, in this study, we learn multifaceted information of media content based on consumer preference from the history of a user's actions with regard to media contents and predict the popularity of contents using multifaceted information. More specifically, we apply Word2vec [13], a popular tool in the field of natural language processing, to the editing history of Wikipedia regarding content and obtain a vector representation of the content. Subsequently, popularity prediction is performed using Wikipedia's number of inbound links as an popularity index of each contents. The overview of this research is shown in Figure 1. The number of inbound links is used to estimate the popularity and importance of blogs [14] and webpages [15]. We also perform qualitative analysis on the vector representation aquired from the editing history.

The contribution of this study is as follows.

- We show that vector representation of media content can be learned from a user's action history on the web.
- We show that prediction accuracy improves as a result of using vector prediction learned from the editing history in media content popularity prediction.
- We analyze the difference between the case in

 ISBN: 978-1-61208-603-3

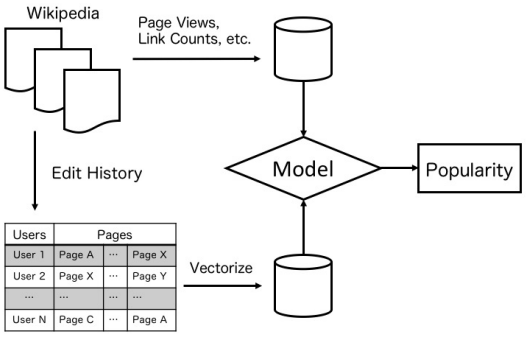

Figure 1. Overview of proposed method.

which prediction accuracy improves and that in which it deteriorates, by performing clustering on contents.

The remainder of the paper is organized as follows. Section 2 discusses related research, and Section 3 explains the proposed method. Section 4 describes preliminary experiments to verify the prerequisites of the proposed method, Section 5 describes experiments to verify the effectiveness of the proposed method, and Section 6 discusses experimental considerations and the development of the method. Section 7 presents the conclusions.

II. Related work

In this section, we describe research related to the proposed method. After discussing research on popularity prediction, we describe research related to Wikipedia editing and the learning of vector representations. Finally, we describe research measuring the popularity of pages from the number of inbound links.

A. Product sales and popularity prediction

Popularity prediction of products and media contents is important in deciding on a marketing strategy [3]. With the spread of the Internet and smartphones, a vast amount of data is available. Thus, many studies on popularity or sales prediction using these data have been conducted. For instance, Choi [16] used volume of queries searched in search engine to predict sales and some economic indicators, and Hozumi [7] used data obtained from various social media to predict popularity.

It has been shown that prediction accuracy can be improved using word-of-mouth information on the problem of sales and popularity prediction of products [9]. For example, Liu [17] attempted to explain the transmission of word-of-mouth information and the revenue of a movie, and Garber [18] predicted the degree of success of new products using word-of-mouth information. In addition, Yu [6] performed sentiment analysis on user information posted on the review site, and then predicted movie revenue.

Considering multifaceted information of products based on consumer preferences is similarly important to consumer word-of-mouth information. Collaborative filtering, a typical method in product recommendation, is a model that reflects user preferences[10][11]. There is a self-reinforcing aspect in the popularity of products, and popularity is known to affect consumer decision-making [19][20]. From this, it seems that consumer preference is relevant to product popularity. However, in the research that predicts the sales and popularity of products, only a few cases took consumer preferences into consideration.

In this study, we aim to conduct popularity prediction considering multifaceted information, such as genre, based on user preference.

B. Studies of Wikipedia

The online encyclopedia Wikipedia is a social medium that is used and edited by many users. Wikipedia covers a wide range of contents, and many studies has been conducted. Specifically, Milne [21] and Strube [22] used Wikipedia as a large corpus of knowledge and its relationships with a link structure as a knowledge base, and Welser [23] and Butler [24] focused on social aspects related to Wikipedia.

Among studies on Wikipedia, there are also studies that focus particularly on editing and editing behaviors. Both registered users and unregistered general users can be Wikipedia editors. Nov [25] examined their motivation by conducting questionnaires for registered users of Wikipedia. In addition, Welser [23] analyzed the social role of registrants' editing behaviors. In the context of popularity prediction, Hozumi [7] used the number of page edits as a prediction feature.

C. Learning multifaceted vector representation

Attention has been paid to a method of acquiring multifaceted information, such as semantic representation in a word with respect to each element, when the series data is given as an input. Word2vec [13], which acquires a distributed representation of words, and [26], which derived from that research and acquires the representation of sentences, has become a major tool. It has been shown that the vector representation learned by Word2vec contains the semantic representation of the word, and the learned vector representation is used for various tasks.

Methods for learning vector representation have also been proposed in fields other than natural language processing, for example, Node2vec [27] takes a graph structure as input and returns a vector representation of each node in the graph. In addition, Barkan [28] learned vector representations related to products from a user's product browsing history and used it for recommendation problems.

D. Link structure as a popularity indicator

The idea of a method for estimating the importance, quality, and popularity of each page from the link structure of a web page has been proposed. PageRank [15], which considers a page with a several links or a page to which a link is affixed from an important page as an important page, and then positions it on a higher level is one typical method. PageRank was introduced to Google's search

 ISBN: 978-1-61208-603-3

engine as an index of the popularity or attention degree of a webpage, with positive results.

Examples of using the link structure as an index of popularity or importance are also available outside webpages. Leskovec [29] showed that link structure in blogs follows a power law, and Wu [14] uses link structure to rank blogs. In addition, Kliegr [30] treated the number of links on each page in Wikipedia as a popularity index.

III. Proposed method

In this section, we explain the proposed method. We obtain vector representations of media contents from Wikipedia editing history. Then, we use the obtained vector representation to predict the popularity of the contents.

A. Learning content vectors

This section describes a method for learning vector representations of contents (content vectors) from the editing history of Wikipedia. Each user is assumed to edit pages based on his or her interests. In addition, contents adjacent to each other in a user's editing sequence are considered to be similar contents based on that user's interests.

In the proposed framework, every content is mapped to a unique vector, represented by a column in a matrix C. The column is indexed by the position of the content in the content space. The editing history for each user is arranged in chronological order, and the concatenation or sum of the vectors is then used as feature for prediction of the next content in an edit sequence. More formally, given a sequence of content edits $c_1, c_2, c_3, ..., c_T$, a vector expression of each content c is learned so that c_t can be predicted by the content $c_{t-k}, ..., c_{t+k}$ existing before and after that. Let C be the matrix that maps each content c to a single vector. Learning of mapping matrix C is performed using continuous bag-of-words (CBOW) or skip-gram [31].

B. Popularity prediction

Popularity prediction is performed by providing model features obtained from Wikipedia and a vector representation of contents, as described in the previous section. The features obtained from Wikipedia, including the numbers of page views, edits, and inbound links, are used.

Multilayer perceptrons (MLPs) are used for monthly features and features from content vectors. Let e_c^t be the monthly number of edits during a month t for a content c, v_c^t be the number of page views, and l_c^t be the number of inbound links. The monthly feature $\boldsymbol{X}_{c,M}^t$ is structured as

$$\boldsymbol{X}_{c,M}^t = [e_c^t, v_c^t, l_c^t] \quad (1)$$

and provided as an input to an MLP for monthly feature, MLP_m.

In addition, let T_c be the vector representation of c obtained from C learned from the content edit history. The MLP for content vector, MLP_c is provided with T_c as an input separately from MLP_m. The outputs obtained from respective MLPs are concatenated and transmitted to a new MLP, MLP_p, as an input to obtain the popularity as the final output.

We provide $\boldsymbol{X}_{c,M}^t$, T_c, and the correct label y (number of inbound links) to the model and jointly train MLP_c, MLP_m and MLP_p.

IV. Preliminary experiment

In this section, we analyze the preconditions of points to be verified when using the proposed method. Experiments are conducted using the Japanese version of Wikipedia, and the popularity prediction targets are media contents such as anime, manga, and games. The premise of popularity prediction using the proposed model is the assumption that user's preferences are reflected in the sequence of Wikipedia's editing history. In addition, preprocessing of obtained data used in experiments is explained. The data used for popularity prediction are obtained during September 2015 to the end of July 2016.

A. Data

The Wikipedia data are acquired from MediaWiki [32]. We collect hourly page views of Japanese Wikipedia pages from dump data. Based on the acquired hourly data, the total value of the page views for 24h is calculated and taken as the data of daily page views.

As for the monthly data, the average value of page views in the previous month and the total number of edits are calculated. In addition, the number of inbound links on the first day of each month is calculated as monthly data. That is, as the value of each feature in October 2015, average page views and total edit counts are calculated based on data from September 1 to 30, 2015, and the number of inbound links for October 2015 is the value for October 1, 2015.

The editing history of the page by the editors is acquired as follows. First, Wikipedia-registered editors who edited Wikipedia pages concerning the anime, manga, and game categories is selected. Then, to secure a sequence length long enough to enable the learning of content vectors, editors with long edit histories are selected. As a result, 2,500 editors corresponding to the the top 5% of the total number of editors are selected and analyzed. The lowest number of editing times by a selected user is 86.

After learning the content vector using the edited sequence of the selected user, future popularity prediction is made using the created monthly data and content vector.

B. Preprocessing

1) Title integration: On the Wikipedia pages of media contents that are serialized for a long time, there are character pages, and related work pages in addition to the main page. Since the contents originally described on one page are distributed and described on multiple pages, the main page of such titles is distributed to related pages, with the result that the number of pages viewed and the number of inbound links are reduced. This also reduces the apparent page views and the number of inbound links. To solve this problem, we use "Path Navi" to describe how the structure of a Wikipedia page relates the page to the main page. If there is a page with "Path Navi" among all pages to be analyzed, then "Path Navi" is analyzed and correspondence with the main page is acquired. When there is "Path Navi," we address the problem that the apparent value decreases by adding the number of edits, page views, and the number of inbound links used for prediction to be added to the value of the main page (in the absence of "Path Navi," this is not done).

 ISBN: 978-1-61208-603-3

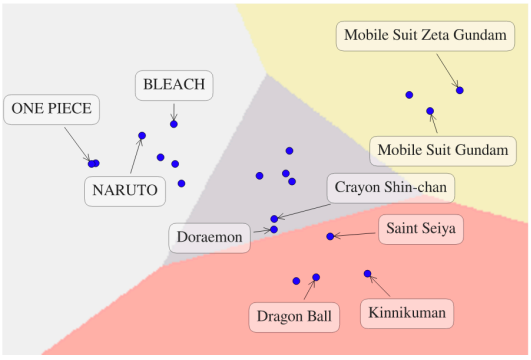

Figure 2. Clustering result of content vector.

2) Selecting target titles: In this study, animation, manga, and games are subjected to popularity prediction. The acquisition of the title to be predicted is carried out in two stages. First, a page listed under the Wikipedia categories "Anime," "Manga," or "Game" other than pages listed under category "List of xx" is obtained. Subsequently, contents are integrated based on "Path Navi," and those with the top 2,000 inbound links are selected. Then, contents that, are not included in a trained vector and whose inbound links shift more than 10% within a month are excluded, because the case in which the number of inbound links fluctuates suddenly is excluded from the prediction by the current model. As a result of narrowing down, 1,547 works are analyzed.

C. Qualitative analysis of Wikipedia editor

In this study, we focus on the top 2,500 editors who registered as editors before October 2015. As a prerequisite for vectorization, the user preferences must be reflected in the editing series, and titles similar in some way must be adjacent in the editing series. Three editors are randomly selected from the 2,500 editors, and a qualitative analysis is performed on a portion of their editing series.

In vectorization, we use the surrounding 10 titles to obtain a vector representation of a content. Accodingly, fragments of the editing history of 10 sequence lengths are selected and analyzed. We randomly specify a position from the edit series of a randomly selected editor, and then select 10 titles to be edited later. Table I shows titles that are included in the selected sequences.

As a result of selecting three series, works related to "Lupin the Third" and those such as "The Kindaichi Case Files" were arranged in series 1. In series 2, gag comics of 2010 such as "One-Punch Man" and "The Disastrous Life of Saiki K." are lined up. In series 3, works related to "JoJo ' s Bizarre Adventure" and many games related to American comics are seen.

In summary, adjacent titles in the editing series tend to have similar genres, authors, fashion ages, and categories as products, and similar common features, reflecting the user's preferences. This result suggests that a vector of target titles can be represented using a vector representation of adjacent titles in an editing sequence.

D. Obtaining a content vector

We describe a method for learning a vector representation of each content from the editing history of a Wikipedia editor. An edit series from editors with several edits of pages belonging to categories targeted for popularity prediction is used. Because it is difficult to learn from an editor's data with few edits, top 5% editors in terms of number of edits is targeted for vectorization. The number of edits by the selected editor is between 86 and 20,793.

The number of epochs is 50,000, the window size is 10, and number of appearances of contents in the editing series is 10 or more in vectorization. CBOW and skip-gram [31] are used as models for vectorization and the size of vector is set to 50.

After learning, qualitative analysis is performed on the obtained vector. To verify the effectiveness of the learned vector, four contents are selected and the top five contents, which are most similar to each content, are acquired. As for the contents related to the query content, contents that are broadcasted or serialized at the same time or contents with similar tendencies are selected. Table II lists the queried contents and top 5 closest contents for each query.

In addition, the results of the kernel principal component analysis on the top 20 inbound link contents' vector are shown in Figure 2. The background color is obtained using k-means clustering. In the area on the left, popular content ranks from content of interest to users in their late teens to age 20, such as "Pokemon", "ONE PIECE", and "NARUTO". In the central area, many contents for children, such as "Crayon Shin-chan" and "Doraemon," are seen. In the upper right area, contents related to "Mobile Suit Gundam", "Mobile Suit Z Gundam", and "Gundam series" are located, and in the lower right area, contents popular in the 1980s and 90s such as "Dragon Ball", "Saint Seiya", and "Dr. Slump" are located. Contents having similar characteristics, such as target age and age of broadcasting, are placed in each area. This result indicates that the vector learned contains information such as the genre of each work and the fashion age.

V. Experiment

In this section, based on the result of the preliminary experiments, popularity prediction of content is performed using a content vector. Moreover, we show that prediction accuracy improves siginificantly as a result of adding a content vector as a feature, compared to the baseline.

A. Popularity prediction

Popularity prediction is performed by a model that takes as input the features and content vectors obtained from Wikipedia, using the number of links of each content work page in Wikipedia as a popularity index. The number of inbound links is used as an index of the degree of importance of pages on the web [15] or an index of the popularity of blogs [14]. Therefore, in this research, based on Kliegr [30], the number of inbound links on each page in Wikipedia is used as an indicator of content popularity.

The prediction model consists of three multilayered perceptrons (MLP_m, MLP_c, MLP_p), given monthly data $\boldsymbol{X}^t_{c,M}$, a content vector T_c, or output of MLP_m and MLP_c

TABLE I. Edit history of content pages (examples).

	editor1	editor2	editor3
1	Lupin the Third	One-Punch Man	Marvel vs Capcom (Game)
2	The Kindaichi Case Files	Chagecha	JoJo's Bizarre Adventure (Game)
3	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	JoJo's Bizarre Adventure (Manga)
4	Lupin the Third (Movie)	Chagecha	Deleted page
5	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	Street Fighter: Sakura Ganbaru!
6	Lupin the Third (Movie)	Blue Exorcist	Oh! Edo Rocket
7	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	Spider-Man
8	Lupin the Third (Movie)	Assassination Classroom	X-Men vs. Street Fighter
9	The Kindaichi Case Files	The Disastrous Life of Saiki K.	Marvel Super Heroes (video game)
10	Lupin the Third (Movie)	NARUTO (Movie)	X-Men: Children of the Atom (video game)

TABLE II. Example of contents vector.

original title	SLAM DUNK	Dragon Ball (Anime)	NARUTO (Manga)	Doraemon
1	Yu Yu Hakusho	Dragon Ball (Manga)	NARUTO (Computer game)	Doraemon (Movie)
2	Touch (manga)	Dragon Ball (Anime; Special)	NARUTO (Movie)	Crayon Shin-chan
3	Dr. Slump	Dragon Ball (Movie)	FAIRY TAIL	Doraemon (Movie)
4	Kimagure Orange Road	Dr. Slump	NARUTO (Computer game)	Doraemon (Movie)
5	I"s	Dragon Ball (Anime; Special)	D.Gray-man	21emon

as an input respectively. As the content vector, we use either T_{sg} which is trained using skip-gram, or T_{cbow} which is trained using CBOW. In addition, a prediction based on a model that does not use a content vector is set as a baseline and compared with the proposed method.

MLP_m has a structure consisting of two layers, with each layer having eight units. MLP_c has two layers, with each layer having 32 units, and a dropout layer inserted between the adjacent layers. The outputs of the two MLPs are combined and provided as input to the MLP layer with 24 units to obtain the final output. ReLU [33] is used for the activation function of the layer except for the final layer, and a linear function is used in the final layer. Only MLP_m and MLP_p are included in the baseline model.

Learning is performed by using back propagation, and optimization is performed using RMSprop. The learning rate is set to 0.0001, the number of epochs is set to 800, and the experiments are conducted using early stopping. Here epoch is number of times that data are passed into the model.

For the content that is analyzed, data from October 2015 to March 2016 are used as training data, and those from April 2016 to July 2016 are used as test data. Next, test data are provided to the model, prediction for three months for each content is performed, and the deviation from the actual value is evaluated using the mean absolute percentage error (MAPE). Since an animation, which is a typical content work, has one occurrence that lasts for three months, we set the prediction period to three months in this study.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_{true} - y_{pred}}{y_{true}} \right| \quad (2)$$

The average of MAPE values for all contents is shown in Table III for the cases with and without a content vector. Random seeds are fixed and five experiments are performed. In the case in which the content vector is used, the average MAPE value is reduced by 2.0% in both T_{cbow} and T_{sg}, compared to the baseline. In addition, the p-value by the t test is 0.004 for T_{cbow} and 0.051 for T_{sg}. When T_{cbow} is used, the value of MAPE significantly decreases at a significance level of 1%. The following are the results obtained by T_{cbow}.

TABLE III. Result of popularity prediction.

# Seed	MAPE ($\times 10^2$)		
	Proposed T_{sg}	Proposed T_{cbow}	Baseline
1	0.153	0.148	0.150
2	0.147	0.148	0.154
3	0.146	0.149	0.150
4	0.147	0.148	0.151
5	0.147	0.147	0.151
Average	0.148	0.148	0.151
p-value (vs Baseline)	0.004	0.051	-

B. Prediction term and prediction accuracy

Next, the relationship between the length of the prediction period and the accuracy of the prediction using the proposed method is investigated. The number of inbound links after zero, one, two, three months is predicted for the model with T_{cbow} and the baseline, and prediction accuracy was calculated. Data from October 2015 to September 2016 was used, of which those from July to September 2016 are used as test data for each prediction. Experiments are conducted five times with different random seeds, and the average value is taken as the final result.

Table IV shows the relationship between the prediction period length and the accuracy of the proposed method. In both the proposed and the baseline methods, a longer prediction period is associated with lower prediction accuracy. In the prediction after zero months, the difference in prediction accuracy between the two methods is small, and the t test reveals no significant difference between the proposed method and the baseline. In prediction after one, two, three months, the p-value decreases as the prediction period increases, but only prediction after three months shows a significant difference at the significance level 1%.

TABLE IV. Prediction period and prediction accuracy.

# seed	MAPE ($\times 10^2$)							
	month 0		month 1		month 2		month 3	
	T_{cbow}	baseline	T_{cbow}	baseline	T_{cbow}	baseline	T_{cbow}	baseline
1	1.31E-02	8.12E-03	7.61E-02	7.78E-02	1.19E-01	1.21E-01	1.48E-01	1.50E-01
2	2.11E-02	2.77E-02	7.71E-02	7.86E-02	1.19E-01	1.22E-01	1.48E-01	1.54E-01
3	1.45E-02	9.96E-03	7.63E-02	8.47E-02	1.19E-01	1.21E-01	1.49E-01	1.50E-01
4	3.93E-02	1.93E-02	8.05E-02	8.00E-02	1.23E-01	1.21E-01	1.48E-01	1.51E-01
5	8.45E-03	1.41E-02	8.03E-02	7.97E-02	1.19E-01	1.30E-01	1.47E-01	1.51E-01
Average MAPE	1.93E-02	1.58E-02	7.81E-02	8.02E-02	1.20E-01	1.23E-01	1.48E-01	1.51E-01
p-value (vs baseline)	0.607	-	0.211	-	0.137	-	0.004	-

TABLE V. MAPE value of each cluster.

cluster id	ratio	# contents	cluster id	ratio	# contents
1	19.73	61	9	4.36	31
2	5.27	318	10	-36.09	22
3	1.00	119	11	3.29	183
4	17.72	22	12	-2.07	31
5	-5.64	120	13	11.96	31
6	18.92	1	14	1.36	246
7	2.66	132	15	4.81	23
8	3.20	192			

C. Prediction accuracy of cluster

As a result of calculating MAPE, there are contents with improved accuracy and contents without, compared with the baseline. To analyze the difference between them, clustering by content vector is performed, and the average MAPE value for each cluster is calculated. After that, the accuracy of prediction for each cluster with or without a content vector is compared. Table V shows the number of contents included in each cluster and the average MAPE improvement over the baseline. Clustering is conducted using k-means, and the number of centroids is set to 15 with respect to the total number of contents of 1,547.

As a result of clustering, the number of contents included in each cluster is between 22 and 318 except for cluster six. Of the clusters consisting of 50 or more contents, we focus on five, cluster 1, 2, 5, 8, and 11, in which the average MAPE value is greatly improved over the baseline, and we examined the top 20 inbound link contents in each cluster.

VI. Discussion

In this section, we discuss experimental and observational results obtained from the editing history analysis, vectorization of contents, popularity prediction, and clustering analysis of popularity prediction.

A. Edit history sequence

As a prerequisite for vectorizing the media contents, it is required that the contents adjacent to each other in the input sequence be similar. Because Wikipedia editors edit pages with which they are familiar, it is expected that adjacent contents in an editing history sequence are similar in some sense. Editors who edit Wikipedia pages related to media content are selected randomly, and a qualitative analysis is performed on a portion of their editing history.

As a result, it becomes clear that adjacent contents are pages relevant to the same contents, contents having the same genre, contents having the same author, or contents related in some other sense. Vector representation of content trained by using an editing series is found to contain information, such as genre, author information, and fashion age.

B. Vectorization

After evaluation of editing history, two qualitative analysises of the content vector are conducted. First, the nearest content of the selected title is retreived, and their similarity is discussed. Second, contents with the most inbound links are selected and their position in the vector space is projected and visualized. As shown in the Table II, the contents located around a query content in the learned vector space have some relevance to query content. The most frequent relationship is that between the targeted content and its related work. In addition, if the authors are the same person (“Dragon Ball” and “Dr. Slump”, “Doraemon” and “21 Emon”) or contents are broadcasted in the same era (“Slam Dunk” and “Yu Yu Hakusho”), they are close in the vector space. There is also a tendency that similar genres or target age contents are located closer in the vector space.

Figure 2 shows a two dimensional projection of vector representations of the top 20 contents in terms of the number of inbound links among the contents targeted. We perform k-means clustering using the values of the first and second components obtained by the kernel principal component analysis. As a result, the first cluster (left) contains popular contents in a wide range of layers, primarily in late teens and 20s men, such as “Pokemon”, “NARUTO”, “ONE PIECE”. The second cluster (center) contains a contents, such as “Crayon Shin-chan” and “Doraemon”, that are familiar to a wide age group, that have persisted for a relatively long time. The third cluster (top right), contains the series related to “Mobile Suit Gundam.” This suggests that the Gundam series has a large distance from other clusters and different user layers. In fact, the “Gundam series” is known to have substantial support from middle-aged and older men. The last cluster (bottom right), contains content, such as “Dragon Ball” and “Saint Seiya”, that prevailed from the 1980s to the early 1990s. It seems that the differences among clusters are the main strata or age of fashion. From the above results, it is considered that the content vector learned using the editing history represents multifaceted information of contents considering the degree of similarity for multiple scales, such as content genre and age, based on the user’s preference.

 ISBN: 978-1-61208-603-3

C. Popularity prediction

For the contents analyzed, the number of inbound links of Wikipedia in three months is predicted. In addition to the monthly input features, content vectors learned using CBOW or skip-gram methods from the editor's history are provided to the model. As a result, prediction accuracy improves by 2.0% in both the CBOW and skip-gram cases as compared to the baseline. This is because the user's preference for the target content is incorporated into the model by inputting the targeted content information as the content vector. Monthly features, such as the number of page views, provided to the model are considered to capture short-term fluctuations. However, those monthly features cannot consider the user's preference for the content formed over the long term. The content vector is trained from the editing history of Wikipedia, and it is considered that the content vector provides information regarding the user's preferences to the contents formed over the long term and information on the user layer that supports the target content, to the prediction model. As a result, it is considered that prediction accuracy is improved by giving the content vector to the model in the popularity prediction.

D. Prediction term and prediction accuracy

We also investigate the relationship between the prediction accuracy improvement and the prediction period using the proposed method. In the prediction after zero months, the model can use the number of inbound links of the current month, which is the prediction target, and hence, no significant difference in MAPE value between the proposed method and the baseline is seen. The accuracy of the proposed method is lower than that of the baseline because the proposed model has more features, which is unnecessary for this case. From the prediction results after one, two and three months, it is clarified that the prediction accuracy declines as the prediction period increases in both the baseline and proposed method. From the results of the statistical significance test between the proposed and baseline methods, the p-value in the statistical test between the two methods clearly decreases with a longer prediction period. When the prediction period is short, prediction based on features of the current month is relatively easy, while when the prediction period becomes longer, fluctuations due to characteristics of the target content tend to become larger, and this makes prediction by the proposed method effective. In this experiment, the proposed method significantly exceeds the baseline method at 1% significance level in the prediction three months after. In addition, with regard to prediction over a period of four or more months, since the animation work that occupies the majority of content subject to this research is broadcast every three months, we do not make predictions because new releases might not be made more than three months before the beginning of the broadcast.

E. Cluster analysis

Content is clustered based on the content vector obtained, and clusters whose prediction accuracy is substantially improved or degraded compared to the baseline are examined. We focus on clusters containing more than 50 contents among clusters with large changes in accuracy and analyze them by checking the top 20 inbound link contents in each cluster. In the case of using a content vector, prediction accuracy improves as the genre of the work and the supporting user layer are strongly included in the content vector. In contrast, it is considered that prediction accuracy decreases when the user layer deployed in a plurality of media and the supporting user layer becomes wider.

VII. Conclusion and future work

In this study, we learned multifaceted information considering multiple degrees of similarity, such as product categories, based on consumer preferences, and made popularity prediction of contents using them. Specifically, we applied a low-dimensional vector representation acquisition method using Word2vec to a user's editing history in Wikipedia. After that, we showed that the accuracy of the popularity prediction improves by providing a learned vector representation to popularity prediction model. The prediction accuracy using the proposed method significantly improved when the prediction period was long. This was probably because if the prediction period is long, not only the current popularity provided as input but also information regarding the content become important. In addition, it is suggested that prediction accuracy becomes higher as the acquired vector representation strongly includes factors such as the genre of the content and the user layer that supports it, based on the comparison result of prediction accuracy for each cluster.

In this experiment, we used the editing history in Wikipedia as the user's action history on the web. However, the proposed method can be applied as long as it can acquire sequence data on objects considered to be of interest to users. Therefore, it can be applied not only to the editing history but also to the action history of a wide range of users, such as the posting history of the review text at a review site and the browsing history of a page at an online shopping site. Thus, a possible extension of this work is to learn vector representation of other user action histories and apply learned vector to other tasks.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP25700032, JP15H05327, JP16H06562.

References

[1] "Anime Industry Data, The Association of Japanese Animations (AJA)," http://aja.gr.jp/english/japan-anime-data, accessed: 2017.07.15.

[2] "Current status of content industry and direction of future development, Ministry of Economy, Trade and Industry," http://www.meti.go.jp/policy/mono_info_service/contents/downloadfiles/shokanjikou.pdf, accessed: 2017.07.15.

[3] R. J. Kuo and K. Xue, "A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights," Decision Support Systems, vol. 24, no. 2, 1998, pp. 105–126.

[4] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment." in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 155–158.

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

 ISBN: 978-1-61208-603-3

[6] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: A case study in the movie domain," IEEE Transactions on Knowledge and Data engineering, vol. 24, no. 4, 2012, pp. 720–734.

[7] J. Hozumi et al., "Consumer trend prediction system using web mining," The Japanese Society for Artificial Intelligence, vol. 29, no. 5, 2014, pp. 449–459.

[8] "Asia Trend Map," http://asiatrendmap.jp/en, accessed: 2017.07.15.

[9] C. Dellarocas, X. M. Zhang, and N. F. Awad, "Exploring the value of online product reviews in forecasting sales: The case of motion pictures," Journal of Interactive marketing, vol. 21, no. 4, 2007, pp. 23–45.

[10] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994, pp. 175–186.

[11] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating "word of mouth"," in Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co., 1995, pp. 210–217.

[12] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in Proceedings of the 23rd international conference on World wide web. ACM, 2014, pp. 867–876.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111–3119.

[14] Y. Wu and B. L. Tseng, "Important weblog identification and hot story summarization." in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 221–227.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

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

[17] Y. Liu, "Word of mouth for movies: Its dynamics and impact on box office revenue," Journal of marketing, vol. 70, no. 3, 2006, pp. 74–89.

[18] T. Garber, J. Goldenberg, B. Libai, and E. Muller, "From density to destiny: Using spatial dimension of sales data for early prediction of new product success," Marketing Science, vol. 23, no. 3, 2004, pp. 419–428.

[19] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," science, vol. 311, no. 5762, 2006, pp. 854–856.

[20] Y. Chen, Q. Wang, and J. Xie, "Online social interactions: A natural experiment on word of mouth versus observational learning," Journal of marketing research, vol. 48, no. 2, 2011, pp. 238–254.

[21] D. Milne and I. H. Witten, "Learning to link with wikipedia," in Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008, pp. 509–518.

[22] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in AAAI, vol. 6, 2006, pp. 1419–1424.

[23] H. T. Welser et al., "Finding social roles in wikipedia," in Proceedings of the 2011 iConference. ACM, 2011, pp. 122–129.

[24] B. Butler, E. Joyce, and J. Pike, "Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia," in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2008, pp. 1101–1110.

[25] O. Nov, "What motivates wikipedians?" Communications of the ACM, vol. 50, no. 11, 2007, pp. 60–64.

[26] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in ICML, vol. 14, 2014, pp. 1188–1196.

[27] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 855–864.

[28] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," arXiv preprint arXiv:1603.04259, 2016.

[29] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in Proceedings of the 2007 SIAM international conference on data mining. SIAM, 2007, pp. 551–556.

[30] T. Kliegr, V. Svátek, K. Chandramouli, J. Nemrava, and E. Izquierdo, "Wikipedia as the premiere source for targeted hypernym discovery," Wikis, Blogs, Bookmarking Tools Mining the Web 2.0, 2008, p. 38.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[32] "MediaWiki," https://dumps.wikimedia.org/, accessed: 2017.07.15.

[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

 ISBN: 978-1-61208-603-3