ICIC Express Letters
Volume **13**, Number **5**, May **2019**

 ISSN 1881-803X
pp. **409–413**

AN EVALUATION OF WEB ARTICLE SIMILARITY BASED ON USER COMMENTS

Katsuyuki Shimoyama, Atsushi Ueno and Tomohito Takubo

Graduate School of Engineering
Osaka City University
3-3-138 Sugimoto Sumiyoshi-ku, Osaka-shi 558-8585, Japan
shimoyama@kdel.info.eng.osaka-cu.ac.jp; { ueno; takubo }@info.eng.osaka-cu.ac.jp

Received October 2018; accepted January 2019

Abstract. *Digital contents include various types such as texts, illustrations, and musics. In the research field aimed at recommending these contents, it is also researched to extracting the feature of the contents from the user comments. However, there are few studies comparing features extracted from the content itself and user comments in terms of similarity evaluation of contents. In this paper, we make a hypothesis that "user comments contain features of contents as much as the contents themselves", and then conduct two experiments using Web articles. The results show the usefulness of using user comments for evaluating the similarity of digital contents.*
Keywords: Document classification, Recommendation, User comments

1. **Introduction.** Digital contents of various forms such as texts, illustrations, and musics are published on the Web. In order to recommend these to users, it is necessary to define features of the contents. When extracting features from contents, there are two main approaches. The first is an approach which extracts features from the contents themselves, such as words in the document, and pixels in the images. This approach is generally known. The other is an approach that focuses on metadata of contents [1, 5, 6, 9]. Contents recommendation has been extensively studied, but there are few studies comparing features extracted from the content itself and metadata [2, 7]. In this research, we focus on the method of extracting features of contents from user comments, and compare these two approaches from the aspect of evaluating the similarity of contents.

In our research, we make a hypothesis that "the user comments contain features representing the digital contents as much as the contents themselves". In order to verify this hypothesis, we conduct two experiments on Web articles and show that the method of extracting the features of contents from user comments is useful in evaluating the similarity between contents.

2. **How to Evaluate Similarity between Web Articles.** In this section, we explain the method of obtaining the distributed representations of the Web articles and the distance function for calculating the similarity between the articles from the distributed representations.

In Hayashi and Onai's research [1], they obtained positive and negative expressions from the review, estimated user preferences, and proposed movie recommendation system. This research does not use reviews for recommendation of the contents themselves. And this research aims to extract another movie review from the given movie review. Therefore, our research is different from Hayashi and Onai's research in the point of that it extracts the contents themselves using user comments.

DOI: 10.24507/icicel.13.05.409

In Li et al.'s research [2], they proposed article-scoring method using both articles and user comments given to them. This research evaluated the usefulness of using comments for the same purpose of our research. Specifically, the topic estimation was conducted and their accuracies were compared, but it did not compare human subjective judgments and the similarities between Web articles calculated by their methods. In our research, we evaluated the usefulness of comments including this point.

2.1. **Learning distributed representation of Web articles.** As an existing method, Term Frequency-Inverse Document Frequency (TF-IDF) and Paragraph Vector (PV) are used in our experiments. In addition, we propose Word Vector-Inverse Document Frequency (WV-IDF) as a new method to obtain distributed representations of documents.

Word Vector (WV) is a method to obtain distributed representations of words [3]. Distributed representations of words acquired with WV have the property that can operate on word vectors such as "king − man + woman = queen", and this calculation result is known for being close to human intuition.

PV extends WV to obtain distributed representations of the documents [4]. PV obtains distributed representation of the document by inputting a vector corresponding to a document in addition to a word sequence.

WV-IDF is obtained by replacing the TF value of TF-IDF with WV. Their distributed representations of documents are expressed as follows.

Definition 2.1.

$$v(d_j) = \frac{\sum_{t_i \in d_j} \vec{wv}(t_i) \cdot \mathrm{idf}_{t_i}}{N_{d_j}}$$

Lemma 2.1.

$$\mathrm{idf}_{t_i} = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

Here, $v(d_j)$ is the WV-IDF vector corresponding to the document d_j, $wv(t_i)$ is the WV vector corresponding to the word t_i, idf_{t_i} is the IDF value of the word t_i, and N_{d_j} is the number of documents that have the word t_i. $|D|$ is the number of documents and $\{d : d \ni t_i\}$ is the number of words in the document. This distributed representation is considered to have properties similar to WV.

2.2. **Calculating similarity between Web articles.** After obtaining the distributed representations, we calculate the similarity between the articles with a distance function.

We use cosine similarity as the distance function for TF-IDF and PV. We use two types of the distance functions for WV-IDF: cosine similarity and Euclid distance. The distributed representation of a document by the WV-IDF corresponds to the weighted average by the IDF value of the words in the document. Therefore, not only cosine similarity but also Euclid distance seems to be suitable as a measure for the similarity between documents.

3. **Experiments.** We verify our hypothesis that "the user comments contain features representing of the digital contents as much as the contents themselves" through two experiments. In particular, we conduct two experiments with Web articles. Through these experiments, we compare the corpora of the articles' body and their user comments in terms of evaluating the similarity between Web articles. In the first experiment, we compare the accuracy of classifying categories of Web articles. In the second experiment, we subjectively evaluate the similarity between Web articles by questionnaire.

3.1. **Experiment setting.** The data set used in the experiment consists of Web articles and their comments crawling from Yahoo! Japan's news site for two days[1]. In this news site, articles and their comments are published in raw text. Therefore, the burden of collecting the text of the article and their comments was small. The condition of articles to be collected is to have at least one comment.

The numbers of articles collected for the experiments are 31 in the category "domestic", 48 in the category "international", 42 in the category "economy", 69 in the category "entertainment", 64 in the category "sports", 12 in the category "IT", 20 in the category "science", and 50 in the category "region" and the sum of all articles is 336. The articles in the category "life" are excluded because most of them are reprinted from other Web services.

We make three data sets from these articles: article body only (Body), user comments only (Comments), article body and user comments (Combined). For these data sets, we perform morphological analysis by MeCab[2] using ver.0.0.5 of the new word dictionary NEologd[3]. In addition, we restrict the part of speech used as a corpus to nouns. As the result, the number of vocabularies of each corpus was Body: 2121 words, Comments: 17519 words, Combined: 18315 words.

In learning the distributed representation, the dimension d of PV and WV was examined one by one from 100 to optimize the dimension d. Of course, there is no guarantee that the optimum dimension d of PV and WV is within this range. However, this range was set in consideration of available calculation resources.

3.2. **Document classification.** Using the distributed representation corresponding to 336 Web articles, we estimated the category of the Web article by K-Nearest Neighbor (K-NN). In category estimation, the parameter K of the K-NN was also examined one by one from 20 to optimize the parameter K. Of course, there is no guarantee that the optimum parameter K of the K-NN is within this range. However, this range was set in consideration of available calculation resources. For each combination of the learning methods of distributed representations and corpus, the highest accuracies are summarized in Table 1 with parameters d, K. As a result, the combination with the highest accuracy was in the case of TF-IDF and Body.

Table 1. The Accuracies of the category estimation

Method	Corpus	d	K	Distance	Accuracy
TF-IDF	Body	none	5	cosine	74.70
TF-IDF	Combined	none	8	cosine	73.81
TF-IDF	Comments	none	16	cosine	71.73
WV-IDF	Combined	40	1, 2	Euclid	66.07
WV-IDF	Comments	60	3	Euclid	65.48
WV-IDF	Combined	20	3, 4	cosine	63.10
WV-IDF	Comments	30	14	cosine	62.80
PV	Combined	20	1, 2	cosine	58.33
PV	Comments	40	1, 2	cosine	56.85
WV-IDF	Body	100	4	cosine	46.43
WV-IDF	Body	10	17	Euclid	30.95
PV	Body	10	14	cosine	21.73

[1]http://news.yahoo.co.jp/ (30, 31st Dec 2015)
[2]http://taku910.github.io/mecab/
[3]https://github.com/neologd/mecab-ipadic-neologd/releases/

3.3. **Questionnaire.** We compare human subjective judgments and the similarities between Web articles calculated from distributed representations by this experiment. The purpose of the questionnaires is to investigate whether the features of the digital contents are contained in a large amount in Body or Comments. Ten people including the students of our university were recruited as subjects, and we conducted two questionnaires to them and got twenty answers from them. In this questionnaire, we used eight methods except for Combined corpus at the parameters with the highest accuracies in the experiments of the category estimation. First of all, we presented a list consisting of the top three articles with high degrees of similarity to a specific article for each method. Then we asked them to choose the top three lists which include the articles with high degrees of similarity to a specific article from the eight lists presented.

The results of the questionnaires are summarized in Table 2 in descending order of the numbers of times judged as the first place, the second place and the third place. As shown in this table, the case of TF-IDF and Comments was most supported by subjects.

Table 2. The results of the questionnaires

Method	Corpus	Distance	judged as First	Second	Third
TF-IDF	Comments	Cosine	15	3	0
WV-IDF	Comments	Cosine	5	5	1
TF-IDF	Body	Cosine	0	6	10
PV	Comments	Cosine	0	3	3
WV-IDF	Body	Euclid	0	2	0
WV-IDF	Comments	Euclid	0	1	2
WV-IDF	Body	Cosine	0	0	4
PV	Body	Cosine	0	0	0

3.4. **Discussion.** First, we consider the results of the category estimation. In TF-IDF, there was no significant difference in the accuracies between the case of Body and Comments corpus. However, the accuracies of Comments corpus greatly exceeded the Body corpus in other methods. From this experiment, we conclude that including more sentences in user comments than in article body is effective in calculating similarity between articles by PV and WV-IDF. Through this experiment, we confirmed that user comments contain information about articles' categories.

Next, we consider the results of questionnaires. The case of Comments corpus was occupied of the top human subjective judgment. From the questionnaires, we concluded that similarity between Web articles calculated from Comments corpus is closer to human judgment than that calculated from Body corpus. Through this experiment, we confirmed that user comments are useful for evaluating similarities between articles.

4. **Conclusions.** In this paper, through two experiments, the hypothesis that "user comments contain features of contents as much as the contents themselves" has been proved in the Web articles. If it becomes possible to calculate the similarity between contents using only user comments, we can calculate the relationship between contents such as texts, illustrations, and musics only from user comments. We would like to verify that user comments are useful for calculating similarity of digital contents other than text through similar experiments.

REFERENCES

[1] T. Hayashi and R. Onai, Movie recommendation system using reviews on Web, *The Japanese Society for Artificial Intelligence*, vol.30, no.1, pp.102-111, 2015 (in Japanese).

[2] Q. Li, J. Wang, Y. P. Chen and Z. Lin, User comments for news recommendation in forum-based social media, *Information Sciences*, vol.180, no.24, pp.4929-4939, 2010.
[3] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, A neural probabilistic language model, *Journal of Machine Learning Research*, vol.3, no.3, pp.1137-1155, 2003.
[4] Q. Le and T. Mikolov, Distributed representations of sentences and documents, *Proc. of the 31st International Conference on Machine Learning*, vol.32, no.2, pp.1188-1196, 2014.
[5] T. Hirayama, T. Yumoto and Y. Takahashi, Extraction and presentation of product reputation by attribute evaluation model, *Proc. of the 3rd Forum on Data Engineering and Information Management*, 2011 (in Japanese).
[6] Y. Nakayama and A. Hujii, Extraction method of evaluation condition for review text, *Proc. of the 19th Annual Meeting of the Association for Natural Language*, pp.248-251, 2013 (in Japanese).
[7] M. Slaney and W. White, Similarity based on rating data, *Proc. of the 8th International Conference on Music Information Retrieval*, pp.479-484, 2007.
[8] J. Kim and M. Marneffe, Deriving adjectival scales from continuous space word representations, *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.1625-1630, 2013.
[9] A. Dai, C. Olah and Q. Le, Document embedding with paragraph vectors, *arXiv preprint arXiv:1507.07998*, 2015.