Topic Models Considering the Meteorological Context and Users' Preference

Taku Ito
Research into Artifacts, Center for Engineering
The University of Tokyo
Chiba, Japan
ito@race.u-tokyo.ac.jp

Yusuke Fukazawa
Research into Artifacts, Center for Engineering
NTT DOCOMO
Kanagawa, Japan
fukazawayu@nttdocomo.com

Dandan Zhu
College of Geophysics and Information Engineering
China University of Petroleum
Beijing, China
chrisdan250@126.com

Jun Ota
Research into Artifacts, Center for Engineering
The University of Tokyo
Chiba, Japan
ota@race.u-tokyo.ac.jp

***Abstract*—In this paper, we focus on the weather and seasonal contexts, and we proposed topic models considering the weather/seasonal contexts and users' preference. We compare two topic models. One is a model not considering users' preference, and the other is a model considering it. By evaluating models from view point of word prediction performance (Perplexity), we discovered the importance of topics associated with users' preference.**

Keywords—topic model; context aware; LDA

I. INTRODUCTION

We focus on the relationship between the weather/seasonal context and the Twitter contents in this research, because we consider that weather condition and the season play an important role in the document generation process when user posts on Twitter. Our research helps to estimate users' preference and trends related to weather and season by analyzing Twitter posts, and the analysis can be applied on product demand prediction or recommendation systems depending on weather condition.

However, not all words and topics are related to the weather or seasonal contexts. There are words and topics which are related to users' personal preference. In this research, we investigate the relationship between Twitter contents, weather/seasonal contexts, and users' preference. In order to investigate the relationship between documents and contexts, many researches have proposed context aware topic models as the extension of LDA [1], such as location [2], time [3], and companion [4]. We propose the model which can divide the words/topics related to the weather/seasonal contexts, and the words/topics related to users' preference. This model can extract words/topics which are related to the weather and seasonal contexts more strongly.

II. PROBLEM STATEMENT

We insist that users' preference is important to consider the weather\seasonal context aware topic models. Therefore, we compare the weather/seasonal contexts aware topic model (WSM) with users' preference and the weather/seasonal contexts aware topic model (UWSM) from view point of the document prediction performance. UWSM considers words and topics related to users' preference by including "user topics" in the model, and WSM does not. Words associated with user topics are related to users' preference rather than the weather or seasonal contexts.

Topic models represent the document generation process by probability model. When the likelihood of words is high, the model can represent the proper document generation process. Therefore, we adopt the likelihood of words as the evaluation value of topic models. If UWSM is the proper document generation model than WSM, the likelihood of words in UWSM is higher than that of WSM.

In this research, the dataset is 928051 Japanese tweets which is posted in Tokyo from May 2011 to December 2011 with geographic information.

III. QUANTITATIVE EVALUATION

To measure the ability of topic models as document generative models, we computed perplexity and compare the resulting values. Perplexity is equivalent to the inverse of the word likelihood [4]. Lower perplexity means that words in documents are not surprising to topic models, so lower perplexity is better. Equation 1 shows the definition of perplexity.

$$\text{Perplexity} = \exp\left(-\frac{1}{\sum_{d=1}^{D} N_d} * \sum_{d=1}^{D} \sum_{i=1}^{N_d} \log(w_{di})\right) \quad (1)$$

In equation 1, w_{di} means likelihood of the ith word in document d. We show the result of comparison of perplexity with UWSM and WSM. We show the result in TABLE V.

TABLE I. COMPARISON OF PERPLEXITY

UWSM	WSM
6760.941	7254.022

From this result, perplexity of UWSM is lower than perplexity of WSM. It means that UWSM can predict words in documents more properly than WSM.

IV. QUALITATIVE EVALUATION

We list the results of learning the relationship between weather context and words. TABLE II shows the results of learning by UWSM, and TABLE III shows the results of learning by WMS. Moreover, we list the results of learning the relationship between users' preference and words. TABLE IV shows the results of learning by UWSM. In this section, we extracted words which are related to food using lexico-syntacninc pattern [5]. The method is to extract nouns posted in the shape of ''eat ……''.

TABLE II. RESULTS OF DISTRIBUTION OF WORD ASSOCATED WITH EACH WEATHER TOPIC BY UWSM

topic	topic11	topic1	topic13
temperature average	10.65	24.39	26.57
temperature standard deviation	1.67	3.60	4.77
associated words	yellowtail	fried rice	apple
	grape fruit	genghis khan	dumpling
	tapioca	hotcake	bacon

TABLE III. RESULTS of DISTRIBUTION OF WORD ASSOCATED WITH EACH WEATHER TOPIC BY WSM

topic	topic9	topic6	topic3
temperature average	11.40	24.55	26.13
temperature standard deviation	3.24	3.65	4.32
associated words	yellowtail	grilled chicken	burger
	cucumber	avocado	noodle
	bread	gummi	saury

TABLE IV. RESULTS of DISTRIBUTION OF WORD ASSOCATED WITH EACH USER TOPIC BY UWSM

topic	topic17	topic9	topic20
associated words	burger	grilled chicken	curry
	water melon	raw fish	noodle
	avocado	banana	meat

Tables show the relationship between each weather topics and words associated with the weather topics. Tables contain the mean, the standard deviation of temperature, and three most major words in the weather topic. For example, TABLE II shows that mean of temperature of topics 11 is 10.65 degrees, and the highest frequently words associated with topic 11 are "yellowtail", "grape fruit", and "tapioca". Since TABLE II is the results of user topics, it lists only associated words.

We discuss the relationship between TABLE III and TABLE IV. In TABLE III, "grilled chicken", "burger", "avocado", and "noodle" are associated with weather topics, but in TABLE IV, these words are associated with user topics. It means that these foods are not related to the weather contexts, but related to users' preference. Next, we discuss the relationship between TABLE II and TABLE IV. Words in TABLE II are related to the weather context, and TABLE IV is related to the users' preferences. These tables contain several fruits such as "grape fruit" and "apple" in TABLE II, and "avocado" and "banana" in TABLE IV. From these results, people get interested in "grape fruit" and "apple" according to the reason of temperature. On the other hand, "avocado" and "banana" are dependent on users' preference rather than the weather context. We can eat "banana" in all seasons now, therefore we assume these results are pertinent. UWSM can classify these words by the relationship between contexts. We consider that UWSM will be useful for product demand prediction or recommendation systems. By qualitative evaluation, we are able to show the effect of adding user topics into topic models.

V. CONCLUSION

Our research purpose is to construct topic models which consider both the weather/seasonal topics and users' preference. We associated the 928051 tweet data with the weather data, and proposed USWM.

About proposed topic models, we evaluated models quantitatively by perplexity which represent the word prediction ability, and discovered that dividing users' preference and weather/seasonal contexts improved the prediction ability of models by 7%. Moreover, we evaluated models qualitatively by comparing words distributions, and discovered that user topics work properly in USWM.

We assume that users' preference and weather/seasonal contexts are not clearly different. Therefore in future, we want to construct models considering combination of users' preference and weather/seasonal contexts.

REFERENCES

[1] D. Blei, Y.Andrew, and M.Jordan: "Latent Dirichlet Allocation", Journal of Machine Learning Research 3, pp.993-1022,2003

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing: "A latent variable model for geographic lexical variation", Proc. of Conference on Empirical Methods on Natural Language Processing, pp.1277-1287, 2010

[3] N. Kawamae: "Trend Analysis Model: Trend Consists of Temporal Words, Topics, and Timestamps", Proc. of the fourth ACM international conference on Web search and data mining, pp.317-326, 2011

[4] Y. Fukazawa and J. Ota: "Companion-aware Topic Model", Proc .of Information Processing Society of Japan Journal, Vol.55, No.1 , pp.413-424, 2014

[5] M. Hearst: "Automatic Acquisition of Hyponyms from Large Text Corpora", Proc. of COLING , pp.539-545, 1992.