Extraction and Categorization of Tip Information from Social Media

Yuki Hattori, Akiyo Nadamoto

***Abstract*—In social media, users post freely many kinds of information which is related to personal behavior, experimentation and their own sentiments on the social media. These information is not written in ordinary web pages, but it is sometimes important information for the users. However, it is difficult to extract such important information from social media, because so much information exists. We propose a method to extract such important information from social media. We call such information "tip information". Our proposed tip information is including a user's experiment and it has common important words. We call the common important words "tip keywords". We first extract user's experience sentences from social media based on experience mining method. Next, we extract sentences which are include tip keywords from them. They become tip information. After extracting tip information, we categorize it according to four categories which are "suggestion and recommendation", "restraint", "briefing", and "possible and impossible". Then we present tip information based on each category.**

***Index Terms*—tip information, social media, experience mining.**

I. Introduction

THE social media such as social networking services(SNSs), blogs, and micro blogs becomes popular. Users post freely many kinds of information which is related to personal behavior, experimentation and their own sentiments on the social media. These information is not written in ordinary web pages, and it is specific to the social media. For example, some festival communities in an SNS present experimental information that is not described on official web pages, such as "When you go to the festival by car, you should exit the expressway one exit early. If you exit at the nearest exit, you will hit a terrible traffic jam." The information is an important for users who are not only social media members, but also outside users. Nevertheless, it is difficult to extract such important information from social media because so much information exists. We designate such important and unique information related to social media as "tip information".

We proposed a method to extract such tip information from social media[1]. The definition of our proposed tip information is (1) the information is credible and important, and (2) a user does not know the information. In this research, as a first step of the research we target on definition (1) the information is credible and important. We consider that information based on a user's experiment is more credible than information that is without a user's experiment. Furthermore, many common important words exist in the information that a user thinks an important. We regard the tip information as including a user's experiment and it has common important words. We designate the common important words as "a tip keywords".

In this paper, we propose improvement of the method, which is to extract tip information. The improvement points are the following.

- Categorizing tip information according to four categories.
- Experimenting to ascertain the usefulness of our proposed method.

In our former proposed method, when the system browses a list of tip information for users, it is difficult for them to understand their desired tip information. Because there are many kinds of tip information. In this paper, after extracting tip information, we propose to categorize the tip information according to four categories which are (1)suggestion and recommendation, (2)restraint, (3)briefing and (4)possible and impossible.

Furthermore, our former experiment is not enough to prove the benefit of our proposed method. Then in this paper, we have new experiment to prove the benefit of our proposed method.

We target an SNS as social media. Our target users are not only SNS members but also outside people. The flow of extracting tip information progresses as follows and as shown in Figure 1:

1) The user inputs a query along with tip information that the user wants to know.
2) The system extracts communities from an SNS and browses the list of communities.
3) The user selects a community from the list.
4) The system extracts comments from the 20 threads of the community.
5) It extracts experience sentences from each comment.
6) It extracts tip information from the experience sentences using a tip keywords dictionary.
7) It categorize the tip information according to four categories.
8) It browses the tip information based on each categories.

In this research, we create tip keywords dictionary in advance.

This paper is organized as follows. Section II offers a discussion of related work. Section III explains tip information extraction. Section IV explains categorizing the tip information. Section V presents a prototype system and results of experiment conducted using our system. Section VI presents the salient conclusions of our study.

Manuscript received January 07, 2013. This work was partially supported by Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (24500134).

Yuki Hattori is with the Konan University Graduate School, 8-9-1 Okamoto, Higashinada-ku, Kobe , Japan, e-mail: (el_clasico_123@yahoo.co.jp).

Akiyo Nadamoto is with the Konan University, 8-9-1 Okamoto, Higashinada-ku, Kobe, Japan, e-mail: (nadamoto@konan-u.ac.jp).

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

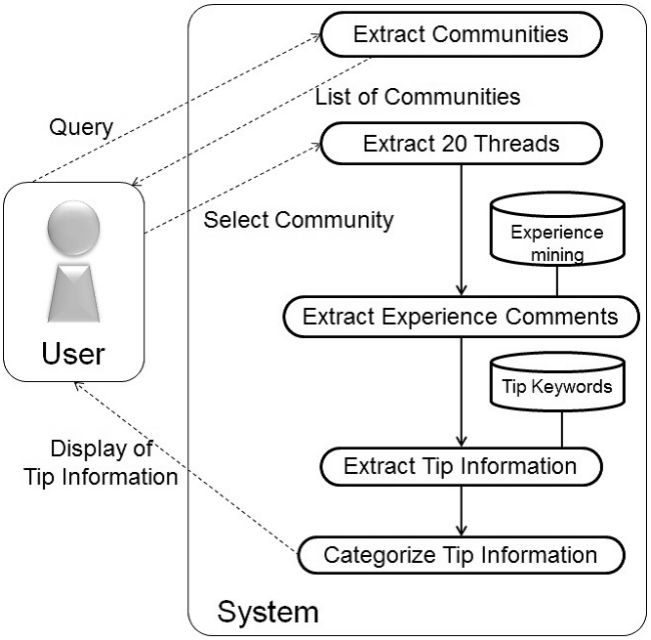

Fig. 1. Flow of extracting tip information

II. Related Work

Recently, social networking has become a popular application on the internet. Therefore, it is important to obtain information from social networking. Ting et al.[2], [3] introduce a method to collect and analyze multi-source social information and, in doing so, to extract social networks from the data.

In the present study, we propose a method of extracting credible and useful information from a social networking service. Inui et al.[4] proposed experience mining, which is aimed at collecting instances of personal experiences automatically along with opinions from an extremely large number of user-generated contents such as weblog and forum posts and storing them in an experience database with semantically rich indices. As described herein, we define credible information as that which has been written based on actual experience. Therefore we extract credible information based on experience mining.

Some studies have extracted useful information as comments written based on the evaluation. Dave et al.[5] develop a method for distinguishing between positive and negative reviews automatically. Liu et al.[6] propose a novel framework for analysis and comparison of consumer opinions of competing products. The system is such that, with a single glance of its visualization, the user can ascertain the strengths and weaknesses of each product in the minds of consumers in terms of various product features. Hu et al.[7] sought to mine and to summarize all customer reviews of a product. This summarization task differs from traditional text summarization. In this approach, one mines only the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. Yu et al.[8] present a model for classifying opinion sentences as positive or negative in terms of the main perspective being expressed in the opinion. Jansen et al.[9] report research results investigating microblogging as a form of electronic word-of-mouth for sharing consumer opinions related to brands. They investigated the overall structure of these microblog posts, the expression types, and the movement in positive or negative sentiment. Morinaga et al.[10] presents a new framework for mining product reputation on the internet. It collects people's opinions related to target products automatically from web pages. Then it uses text mining techniques to ascertain reputations of those products. For this study, we define useful information as comments that are written based on personal experience, this information has not been granted a reputation identifying it as either positive or negative.

We think that useful information contains a common keyword. Turney [11] has been using a semantic phrase in the review. This approach presents a simple unsupervised learning algorithm for classifying reviews as recommended or not recommended. In this approach, the classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. The semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word ”excellent” minus the mutual information between the given phrase and the word ”poor”.

III. Extracting Tip Information

As described in this paper, we propose a means of extracting tip information that is credible and important information. We consider that credible tip information is based on the author's experience, and that important tip information uses some tip keywords. Then we propose a means to extract experience sentences and tip keywords.

A. Extracting Experience Sentences

We extract experience sentences based on experience mining method[4]. Experience mining is intended for the automatic collection of instances of personal experiences as well as opinions from social media. Experience mining consists of a topic object, experiencer, event expression, event type, factuality, and a source pointer. We assume that tip information is based on the author's event expression. We specifically examine only the event expression in experience mining. Their proposed event expression consists of sentiment, happening, and action. They describe the sentiment as predicative expressions of an emotion or subjective evaluation, a happening as predicative expressions referring to a non-volitional event or state which is related to the use of a topic object and has a sentiment orientation, and the action as predicative expressions referring to the experiencers' volitional actions related to the use of a topic object. Then they propose the Japanese Polar Phrase Dictionary[12][13]. It consists of happening words, which consist of verb, adjective, and noun, and their Positive/Negative value as a sentiment of the word. We use the Japanese Polar Phrase Dictionary to extract sentiment and happening sentences.

On the other hand, the action depends on the domain. For example, one action related to shopping is “buy”. One action related to food is “eat”. Then we create an action words dictionary of the domain dependence from each community of the SNSs. Table I presents an example of festival domain action words. We first gather 3000 comments from festival communities. Then we extract the action words manually and

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
EXAMPLE OF FESTIVAL DOMAIN ACTION WORDS.

Verb (in Japanese)		Noun (in Japanese)	
Go (iku)	Buy (kau)	Use (riyou)	Participate (sanka)
Lose way (mayou)	View (miru)	Move (idou)	Excitement (koufun)
Able to (dekiru)	Drink (nomu)	Cheers (kanpai)	Activity (katudou)
Dance (odoru)	Eat (taberu)	Ensure (kakuho)	Discovery (hakken)
Make noise (sawagu)	Wear (kiru)	Shooting (satsuei)	Break (kyuukei)
Listen (kiku)	Search (sagasu)	Action (koudou)	Assembly (shuugou)

TABLE II
NUMBER OF TIP INFORMATION COMMENTS

Communities(Number of comments)	A	B	C	D	E
PL Fireworks Festival(596)	114	155	321	138	98
Autumn Leaves in Kyoto(287)	27	102	137	69	49
a-nation(881)	16	82	338	200	28
Beach in Kansai Area(99)	23	33	11	16	24
Kyoto Gion Festival(203)	42	33	19	24	40

TABLE III
NUMBER OF COMMENTS BY THE TOTAL NUMBER OF PARTICIPANTS.

Total number of participants	PL fireworks show	Autumn leaves in kyoto	a-nation	Beach in kansai area	Kyoto gion festival
5	48	20	7	4	3
4	54	24	14	5	8
3	44	26	38	9	11
2	47	31	151	12	24
1	134	46	149	15	32

TABLE IV
EXAMPLES OF TIP KEYWORDS.

Tip keywords (in Japanese)		
recommendation (osusume)	.. is better (no houga yoi)	How about ..? (ikaga)
should (subeki)	By all means (zehi douzo)	highly recommended (itioshi)
your chance(neraime)	a well-kept secret place (anaba)	convenience (benri)
almost empty (garagara)	a traffic jam (juutai)	congestion (konzatu)

count the term frequency (TF) of the nouns and verbs. We then infer the top 50 words of TF as festival domain action words.

After we create an action words dictionary, we extract experience sentences using a Japanese Polar Phrase Dictionary and an action words dictionary. The sentences become tip information candidates.

B. Creating Tip Keywords Dictionary

We extract tip keywords using our experiment. First, we gathered tip information from comments of SNSs, and extracted common keywords from it. Specifically in our experiment, we targeted festival communities. In our experiment, five participants read 2000 comments and judged the comments as tip information or not. We regard information that is judged as tip information from four participants in five participants as tip information. Then we extract tip keywords from the sentences that were judged as tip information. Tables II and III show the numbers of comments that participants judged as tip information. Table IV presents examples of tip keywords in festival communities.

After we create a tip keywords dictionary, we extract important tip information from the tip information candidates extracted in section III.A.

IV. CATEGORIZING TIP INFORMATION

When we extract tip information, there are so many kinds of tip information. In this case, it is difficult for users to extract their desired tip information. Then we propose we categorize tip information according to four categories. We analyze extracted tip information, we can divide into four categories of tip information (1)suggestion and recommendation, (2)restraint, (3)briefing and (4)possible and impossible. These categories are include each specific keywords. We call the specific keywords “category keywords”. We use category keywords which are specified categories to categorize the tip information. When a comment include a category keyword, it is categorize the category which is include category keyword. In this time, there are multiple category keywords in a comment, we determine multiple categories for a comment.

We extract category keywords using our experiment. In our experiment, the three participants read the comments of all eight themes and judged the comments as tip information or not. After extract tip information, they categorize them according to four categories. The eight themes of communities were the following.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE V
CATEGORY OF TIP INFORMATION

Category	Example of category keywords
suggestion and recommendation	better than, recomend, should
restraint	be discouraged from, should give up
briefing	crowd, at one's best, sold out
possible and impossible	can, can't

TABLE VI
THE NUMBER OF TIP INFORMATION IN EACH CATEGORY

Category	number of comments	rate
suggestion and recommendation	772	33%
restraint	90	4%
briefing	1071	46%
possible and impossible	389	17%

- PL Fireworks Show: This is the largest fireworks festival in western Japan. Every year, many people come to the festival, creating terrible traffic jams.
- Autumn Leaves in Kyoto: In autumn in Kyoto, many tree leaves turn red or yellow. Many people come to Kyoto to see them.
- Beach in Kansai area: This beach is in western Japan.
- Nebuta Festival: In Tohoku area, it is famous traditional festival.
- Nabana no sato: This illuminated park is a famous tourist destination.
- Sapporo Snow Festival: It is the biggest snow festival in Japan.
- Gathering of clams: In Japan, many people come to the seaside to gather clams.
- Tokyo Game Show: It is the biggest game show in Japan.

We extract category keywords from the comments which are categorized four categories by hands. Table V shows example of category keywords in each category. Table VI shows number of comments which is categorized four categories by three participants in our pre-experiment. In this results, briefing category is the most common tip information.

V. PROTOTYPE SYSTEM AND EXPERIMENT

A. Prototype System

We develop our prototype system. For it, we use Ruby[1], with PHP as the programming language. The target SNS is mixi[2] which is the most popular SNS in Japan. In our prototype system, the first user inputs a query related to the theme of a community of which the user wants to know tip information using our system (Figure 2(a)). Then, the system outputs a list of communities (Figure 2(b)). Next, a user selects the community. Then the system extracts comments that include tip information from the community. The system categorizes the comments according to the four categories, and displays in each category(Figure 2(c)).

B. Experiment

We conducted experiment to confirm the accuracy of our proposed method using our prototype system.

[1]Ruby http://www.ruby-lang.org/
[2]mixi http://mixi.jp

Experiment condition

The participants are three. Datasets are above mentioned eight themes which are "PL Fireworks Show", "Autumn Leaves in Kyoto", "Beach in Kansai area", "Nebuta Festival", "Nabana no sato", "Sapporo Snow Festival", "Gathering of clams", and "Tokyo Game Show". The three participants read the comments of all eight themes and judged the comments as tip information or not. The correct answers those judged by two or three of the participants as correct. We calculate precision, recall, and F-measure.

Results and Discussion

Table VII shows the results of the experiment. The average precision is 62% and the average recall is 37%. "PL Fireworks Show", "Autumn Leaves in Kyoto", "Beach in Kansai area", "Nabana no sato", "Sapporo Snow Festival", and "Gathering of clams" are good result whose precision is greater than 60%. However, "Nebuta Festival" and "Tokyo Game Show" are less than 50%. In some unfortunate cases, there are some promotion for shops which is include some tip keywords. However, participants judge the comments is not a tip information but just a shop's promotion. On the other hand, the results of recall are bad. The reason is that some comments are include tip keyword but not include experiment. In this time, participants judged the comments are tip information, but the system judged they are not tip information because the comment is not include experiment sentence.

Moreover, the results differ among participants. Table VIII shows the results of the experiment by each participant. In this way, judgments related to tip information depend on the person making them, which is a point that must be considered in the near future.

VI. CONCLUSION

In this paper, we proposed the method for extracting tip information that is credible and important information from social media. We consider that credible information is based on the author's experience. And we consider that important information uses tip keywords. Then we proposed means to extract comments that consists of the author's experience and tip keywords. After we extract tip information, we categorize it according to four types which are "suggestion and recommendation", "restraint", "briefing", and "possible and impossible". Then we present tip information based on each category. We conducted an experiment demonstrating the accuracy of our proposed methods.

Future work will include the following tasks.

- There are some shop's promotion in the tip information, we will consider how to except such information form the tip information.
- We should consider personalization because different tip information by the users.
- We will categorize tip information based on positive/negative.

REFERENCES

[1] Y. Hattori and A. Nadamoto. Extracting Tip Information from Social Media. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services*, pages 205–212, 2012.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

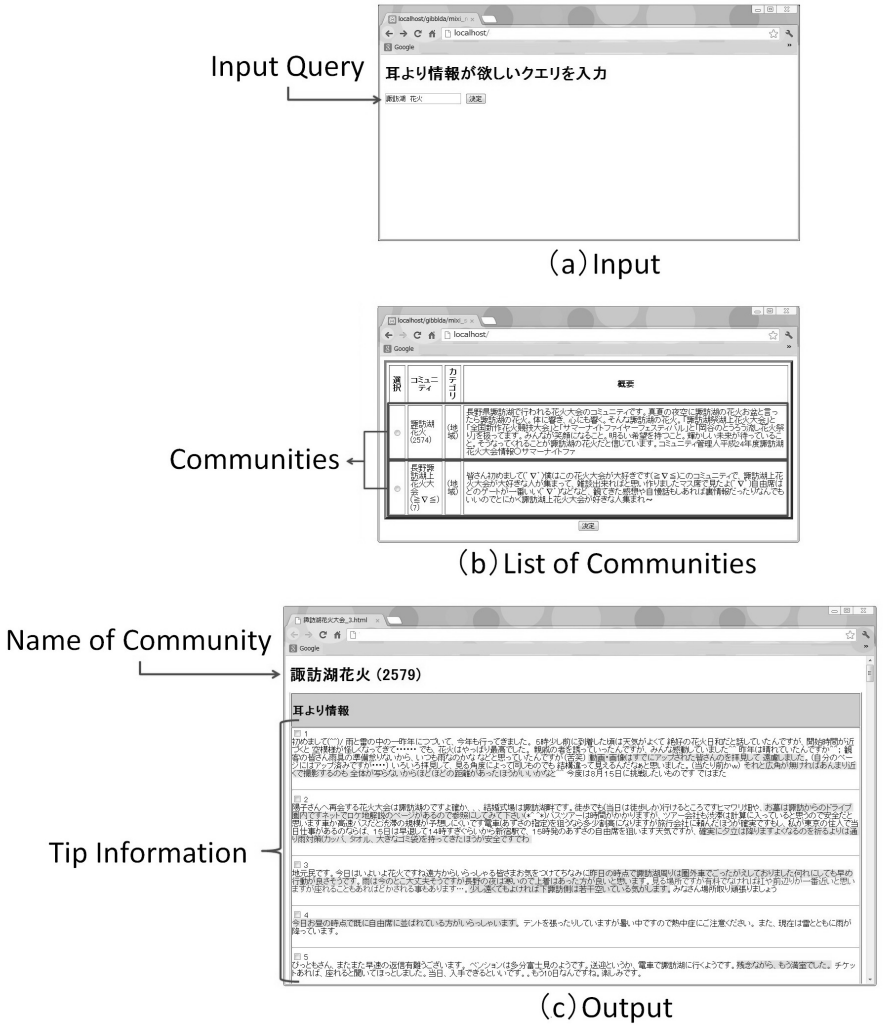

Fig. 2. Display of the result in our prototype system

TABLE VII
ACCURACY OF PROPOSED METHODS

Community	Number of all comments	Number of comments include tip information	Precision	Recall	F-measure
PL fireworks show	867	149	68%	47%	0.56
Autumn leaves in Kyoto	323	44	70%	45%	0.55
Beach in Kansai area	328	40	60%	39%	0.48
Nebuta festival	759	56	41%	34%	0.37
Nabana no sato	290	39	72%	43%	0.54
Sapporo snow festival	183	26	73%	37%	0.49
Gathering of clams	418	37	68%	25%	0.36
Tokyo Game Show	1025	87	45%	27%	0.33
Average	–	–	62%	37%	0.46

[2] I. H. Ting. Web mining techniques for on-line social networks analysis. In *Proceedings of the 5th International Conference on Service Systems and Service Management, Melbourne, Australia*, pages 696–700, 2008.

[3] I. H. Ting, H. J. Wu, and P. S. Chang. Analyzing multi-source social data for extracting and mining social networks. In *Proceedings of International Conference on Computational Science and Engineering*, pages 815–820, 2009.

[4] K. Inui, S. Abe, H. Morita, M. Eguchi, A. Sumida, C. Sao, K. Hara, K. Murakami, and S. Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Proceedins of 49the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 314–321, 2008.

[5] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference*, 2003.

[6] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, pages 342–351, 2005.

[7] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of ACM-KDD*, pages 168–177, 2004.

[8] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of Japanese Society for Artificial Intelligence*.

[9] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power:tweets as electronic word of mouth. In *booktitle of the American Society for Information Science and Technology*, 2009.

[10] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of the 8th ACM*

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE VIII
ACCURACY OF PROPOSED METHODS IN EACH PERTICIPENT

	A			B			C		
Community	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
PL fireworks show	74%	43%	0.54	36%	52%	0.43	72%	39%	0.51
Autumn leaves in Kyoto	70%	38%	0.5	66%	43%	0.52	48%	46%	0.47
Beach in Kansai area	68%	33%	0.44	48%	44%	0.46	58%	41%	0.48
Nebuta festival	46%	26%	0.33	16%	50%	0.24	57%	24%	0.34
Nabana no sato	77%	38%	0.51	31%	67%	0.42	69%	37%	0.48
Sapporo snow festival	81%	36%	0.49	42%	41%	0.42	69%	31%	0.43
Gathering of clams	70%	23%	0.35	14%	16%	0.15	76%	20%	0.32
Tokyo game show	46%	23%	0.31	21%	29%	0.24	53%	20%	0.29
Average	59%	30%	0.39	31%	41%	0.33	56%	30%	0.38

SIGKDD Conference, 2002.

[11] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.

[12] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. Collecting evaluative expressions for opinion extraction (in japanese). In *booktitle of natural language processing*, pages 203–222, 2005.

[13] M. Higashiyama, K. Inui, and Y. Matsumoto. Acquiring noun polarity knowledge using selectional preferences (in japanese). In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587, 2008.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)