Quick Evaluation of Research Impacts at Conferences using SNS

Satoshi Fukuda, Hikaru Nakahashi, Hidetsugu Nanba, Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
{fukuda, nakahashi, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

***Abstract*— We are investigating ways of evaluating research impact as soon as possible after publication. Traditionally, the research impact or importance of academic journals has been evaluated using citation relations, such as the impact factor and the citation half-life. However, these citation-based methods require long periods to evaluate research impact and therefore are not suitable for evaluating the current impact of research papers at conferences. To solve this problem, we are studying the automatic evaluation of research impact using Twitter. Researchers participating in academic conferences often post their opinions or comments on Twitter. Here, research papers (presentations) that have many comments are considered to be outstanding and to have strong impact during the conference. In this paper, we propose a method for automatically aligning tweets with research papers. The procedure consists of the following three steps: (1) detecting valuable tweets, (2) aligning each valuable tweet with a research paper, and (3) calculating the research impact of each research paper by the number of aligned tweets. We conducted some experiments to confirm the effectiveness of our method. From the results, we obtained an MRR score of 0.223, which outperformed a baseline method.**

Keywords—component; Twitter; research paper; research impact

I. INTRODUCTION

Traditionally, the research impact or importance of academic journals has been evaluated using citation relations, such as the impact factor [1]. However, evaluation by these citation-based methods requires long periods. For example, for a given year, the impact factor calculates the impact of a journal using the number of citations during the two preceding years. Therefore, the impact factor does not reflect the latest research impact. To solve this problem, we have been studying the quick evaluation of research impact using Social Networking Services (SNS).

Nowadays, researchers participating in academic conferences often post their opinions, questions, or comments on Twitter. In a tweet, a presenter or some other participant posts an opinion or the answer to a question. Here, research papers (presentations) that have many comments are considered to be outstanding and to have a strong impact in the conference. Therefore, if we can collect tweets for a conference, and align each tweet with a research paper in the conference, we can find high-impact papers in the conference very quickly.

Aligning tweets with research papers provides another benefit for researchers who cannot attend the conference, because they can find various opinions or comments for each paper, which may help them to understand the paper better. Therefore, we propose a method of automatic alignment of tweets with research papers.

II. RELATED WORK

Several measures for evaluating the importance of research using citation relations have been proposed. For example, Hirsch [2] proposed an *h*-index that quantifies the research output of an individual researcher. This index is a simple yet robust citation index and a metric for computing impact factor as a single number. To derive this number, all research papers are sorted in descending order of citation count, and then the highest index (*h*) of a research paper with a citation count of at least *h* is constructed. The *h*-index can be used to evaluate the impact factor of any aggregate of research papers. In addition to the impact factor mentioned in above, Garfield [1] proposed a citation half-life. This measure evaluates the impact of each academic journal by the median age of research papers that were cited by the papers published in the journal. However, these approaches require long periods to evaluate.

Several studies have attempted to solve this problem. One of the earliest research projects was KDD CUP 2003[1]. In this project, the following two tasks were assigned.

1. Citation prediction task: Predict changes in the number of citations to frequently cited papers over time.
2. Download estimation task: Estimate the number of downloads of a paper in its first two months in the arXiv[2].

By estimating citations or downloads, we can expect to evaluate research impacts more quickly. However, the developer of the top system in the download estimation task reported that the results were disappointingly inaccurate[3]. More recently, Yogatama et al. tried the same task with a different dataset [3], but the results were still not encouraging. On the other hand, in the citation prediction task, many researchers have studied the problem of Citation Count Prediction (CCP) for a research paper [4, 5, 6].

[1] http://www.cs.cornell.edu/projects/kddcup/
[2] http://arxiv.org
[3] http://www.cs.cornell.edu/projects/kddcup/download/KDDCup-Task3.ppt

As in other approaches for evaluating research impacts, Vaughan et al. [7] proposed a method that uses web citations, which are citations from the literature on the web. The basic idea is to use web citations instead of citations in journals between research papers, and Vaughan reported that web citations and ISI citations[4] have strong correlations. Sayyadi and Getoor [8] proposed a PageRank based on method [9], which gives an expected future PageRank score using citations that may be obtained in the future. They also combine the citation network, the authorship network and the publication time to rank the future citations of research papers. They called this approach "FutureRank".

Recently, several systems have been developed to evaluate research impacts using SNS. Eysenbach et al. [10] analyzed the prediction of citations of a research paper based on tweets posted on Twitter. They found that Twitter can find the latest academic research paper, and the number of tweets related to the research paper is an important evaluation measure. Weller et al. [11] proposed a method that collects tweets posted during an academic conference and then analyzes citations of research papers using the number of tweets that contain the URL for the academic paper and the number of retweets. Priem et al. [12] constructed a system that measures the evaluation of a research paper based on the number of messages on several social media platforms such as Twitter and Facebook. They called their system "Altmetrics" and described new metrics for measuring in real time.

These studies used only numbers of tweets and ignored their content. In our work, we analyze the comments on presented research papers and consider whether tweets are valuable as evaluation for research papers.

III. EVALUATING RESEARCH IMPACT BY ALIGNING VALUABLE TWEETS WITH RESEARCH PAPERS

In this paper, we propose methods for automatically aligning tweets with research papers and determining whether each tweet is valuable for evaluating the corresponding research paper. We explain our methods for detection of valuable tweets, for alignment of tweets with research papers, and for calculation of research impacts.

A. Detection of Valuable Tweets

In this section, we describe our method for automatically deciding whether a tweet is valuable for a research paper.

Definition of Valuable Tweets

Examples of valuable and valueless tweets are shown in Figure 1[5]. We regard tweets that contain sentiments (tweet 1), questions for a researcher (tweet 2), replies to a question (tweet 3), and comments (tweet 4) as valuable. We also consider tweets containing related URLs as valuable, because these URLs are valuable links, such as related research papers or web sites. On the other hand, we regard tweets that are not related to the research paper (tweet 5), that contain only bibliographic information of the paper, are live tweets (broadcasting) (tweet 6), and retweets (tweets beginning with "RT") (tweet 7) as valueless.

(Examples of Valuable Tweets)
(1) Relationship of onomatopoeia and font is interesting.
(2) Are terms like "loose" and "slowly" also onomatopoeia?
(3) @hijip All onomatopoeia are OK, because the entered onomatopoeia is decomposed into vowels and consonants, and then quantified.
(4) I wonder whether some words such as "loose" and "slow" are characteristic ones that are included in the evaluation of hotel.
(Examples of Valueless Tweets)
(5) I came to the shikakeology session by a 15-minute walk (> <).
(6) "Automatic Organization of Travel Information", they extract some links that are contained from travel blog entries.
(7) RT @nanaya_sac whether the recipe book of packed onomatopoeia is heresy.

Figure 1. Examples of valuable and valueless tweets

Strategies for Detecting Valuable Tweets

We detect valuable tweets using machine learning. For the machine-learning method, we opted for a Support Vector Machine. This method identifies the class of each word. We used the following cue phrases as features for machine learning. Note that the words (cue phrases) used were nouns, verbs, and adjectives. We also used MeCab[6], a Japanese morphological analysis tool to identify the parts of speech.

- **Sentiment lexicons for sentiment analysis**: We use "whether a word in each tweet is contained in an automatically constructed sentiment lexicon [13]" as a feature. This lexicon contains approximately 10,000 Japanese polar phrases with their polarity values.
- **Information gain (IG)**: In variable tweets, particular expressions such as "interesting" or "great" appear frequently. To collect such clue expressions, we employed the information gain (IG) method, which reduces the cost of collecting cue phrases and collects useful words as features. From the result of our pilot study, we obtained 100 cue phrases, which have high IG values.
- **Similarity between a tweet and a research paper**: Character strings similar to valueless live tweets, such as tweet 7 in Figure 1, often appear in research papers. We therefore calculate the similarity between a tweet and a research paper, and use it as a feature. Our similarity measure was used dynamic programming (DP) matching [14], which is useful for measuring the similarity between a tweet and a research paper with their greatly different numbers of characters.
- **Sentence-final expressions**: To identify opinion tweets, such as tweet 4, we used 347 sentence-final expressions, because valuable clues, such as

[4] Institute for Science Information, the world's largest science citation index, provided by Thomson Reuters.
[5] The tweets were written in Japanese, but we translated in English.
[6] http://mecab.sourceforge.net/

modality, tense, and aspect, tend to appear in Japanese sentences.

In addition to the cue phrases above, we use the presence or absence of a question mark and reply to a tweet, which begins with "@", as features for identifying valuable tweets, such as tweets 2 and 3 in Figure 1.

B. Alignment of Tweets with Research Papers

The procedure for the alignment is as follows.

1. Obtain candidate research papers corresponding to the tweet using the time of posting of each tweet and the time of presentation of research papers in the conference. Here, we refer to research papers presented up to 30 minutes before the posted tweet, because participants in academic conferences usually post some opinions and questions concerning the research paper during or after the presentation.
2. Calculate the similarity between a tweet and a candidate research paper using a similarity measure.
3. Align the tweet with the research paper that has the highest similarity.

In aligning tweets with research papers, we take account of the following two points.

- **User information (User)**: The maximum length of a tweet is 140 characters, and there may not be enough information in a tweet to align with a research paper. Therefore, we use all tweets posted by the same user within 20 minutes before and after the target tweet.
- **Sections in each research paper (Sec)**: Tweets do not always mention the whole research paper, but rather mention a part of the paper, such as the methodology or results. Therefore, we use the following two methods in Step 2.

1. Calculate the similarity between a tweet and a section in a research paper.
2. Calculate the similarity between a tweet and a complete research paper.

C. Calculation of Research Impact of Each Research Paper

Finally, we calculate the research impact of each research paper as follows.

(Step-1) Count the number of automatically aligned tweets for each research paper using our method described in Section III-B.
(Step-2) Rank research papers by the number of aligned tweets.

IV. EXPERIMENTS

A. Data Sets and Evaluations

We collected tweets about research presentations using hash tags from Togetter[7]. Our collected tweet data were assigned hash tags that were constructed using "#" and alphanumeric strings. Each conference has a specific hash tag, and participants at the conference contributed tweets with that hash tag. We selected 13 hash tags corresponding to conferences, and collected 4,693 tweets in total. We also collected 291 research papers in total that were published in the conferences. Details of our tweet data and research paper data are shown in Table I. One annotator then manually judged whether each tweet was valuable or valueless, using the criteria in Section III-A. We identified 840 tweets as valuable among the 4,693 found. As the evaluation measures, we used precision, recall, and F-measure.

TABLE I. EXPERIMENTAL RESULTS FOR CALCULATING RESEARCH IMPACTS OF EACH RESEARCH PAPER

Conference name	Hash tag	No. of tweets	No. of research papers
The Japanese Society for Artificial Intelligence	#jsai2010	458	17
	#jsai2012	600	51
	#jsai2014	246	4
Rakuten Research and Development Symposium	#rrds3	706	7
The Association for Natural Language Processing	#nlp2012	268	29
GIS Association of Japan	#gisa2011	296	47
Web Intelligence and Interaction (1) 1st conference (2) 2nd conference (3) 3rd conference (4) 4th conference	(1) #sigwi2	195	11
	(2) #sigwi2	241	16
	(3) #sigwi2	75	8
	(4) #sigwi2	436	11
Data Engineering and Information Management	#DEIM2012	754	63
	#DEIM2013	295	15
	#DEIM2014	155	9

B. Experiment: Detection of Valuable Tweets

Experimental Settings

We explain our method for automatically classifying tweets for a research paper. We used TinySVM[8] as the machine-learning package and used a liner kernel. We performed a two-validation test. We conducted tests using our method and a baseline method. Note that we used nouns, verbs, and adjectives as features.

- **Our method**: Use cue phrases, such as those in Section III-A, and frequency of occurrence of all words, as features for machine learning.
- **Baseline**: Use frequency of occurrence of all words as features for machine learning.

Results and Discussion

The experimental results are shown in Table II. Our method obtained a higher F-measure score than the baseline method. We also compared this method with baseline method by the McNemar test, and a significance level of 0.01 was obtained. This result indicates that using cue phrases described in Section III-A can be useful.

[7] http://togetter.com

[8] http://chasen.org/~taku/software/TinySVM/

TABLE II. EXPERIMENTAL RESULTS FOR CLASSIFICATION OF TWEETS

	Precision	Recall	F-measure
Our method	**0.588**	**0.591**	**0.589**
Baseline	0.581	0.534	0.557

A typical reason for the low precision of our method is that some valueless cue phrases are contained in the cue phrase list constructed by the IG method. For example, "venue (会場)" (which occurs in the cue phrase list) is not useful for identifying valuable tweets because this word does not represent assessment or opinion. To remove such words, we must construct a list of unnecessary words in advance. To improve the recall of our method, we must consider the lack of cues. For classification of tweets, we used the lexicon constructed by Kaji and Kitsuregawa [13]. Assessment words, such as "interesting (面白い)" and "great (素晴らしい)", are contained in this lexicon. However, these words have orthographic variants such as "great (スバラシイ)" and "great (すばらしい)", which are not in the lexicon.

C. Experiment: Alignment of Tweets with Research Papers

Experimental Settings

In this experiment, we used 291 research papers and 840 tweets after removing the valueless ones because those tweets were not related to the research papers. We examined various similarity measures, such as DP matching, cosine similarity, and ROUGE-N [15], and found that DP matching obtained the best performance among them. In the following, we show some tests using our four methods by DP matching and a baseline method.

Our methods

- **DP**: Use DP matching as a similarity measure.
- **DP + User**: Use DP with the same user's tweets within 20 minutes before and after the target tweet.
- **DP + Sec**: Use DP to calculate a similarity with a section in a research paper.
- **DP + User + Sec**: Use DP + User to calculate a similarity with a section in a research paper.

Baseline method

- **Baseline**: All research papers within 20 minutes before and after the target tweet.

Results and Discussion

The recall, precision, and F-measure for our methods and baseline methods are shown in Table III. This table shows that our methods improved recall scores compared with the baseline method. In particular, the DP + Sec method significantly improved both recall and precision scores, and obtained the highest F-measure score. We also compared DP + Sec method with the baseline method by the McNemar test, and a significance level of 0.01 was obtained.

We checked the errors affecting the DP matching method. We found misalignments between research papers presented in the same session in an academic conference and tweets. Figure 2 shows an example. Both research papers and the tweet in this figure are concerned with onomatopoeia. The contents of these research papers are very similar, and there are many words that are included in both of them, such as "onomatopoeia." On the other hand, the important words in the tweet are only "onomatopoeia" and "expression dictionary." As a result, our method could not distinguish the two research papers by analyzing the tweet and incorrectly aligned the tweet with the second research paper. To solve this problem, we focus on assessment expressions in tweets. For example, a tweet posted for a research paper during presentation tends to contain assessment words such as "interesting (面白い)" and "great (素晴らしい)". By analyzing these expressions and taking into account the post and presentation times, we consider that it is possible to identify whether the tweet really was posted during the research presentation.

TABLE III. EXPERIMENTAL RESULTS FOR ALIGNMENT OF TWEETS WITH RESEARCH PAPERS

	Precision	Recall	F-measure
DP	0.492	0.453	0.472
DP + User	0.477	0.438	0.456
DP + Sec	**0.525**	**0.483**	**0.503**
DP + User + Sec	0.514	0.473	0.493
Baseline	0.463	0.370	0.411

[Tweet]
Does an Onomatopoeia expression dictionary exist?
[Paper 1](Correct)
In this paper, we propose a method of automatic extraction of information from a review that uses the features of onomatopoeia, and tag information attached to the customer reviews.
[Paper 2]
Figure 1 is a matrix of the results of the analysis of feature words of onomatopoeia as "ramen" and "noodles".

Figure 1. Example of a failure in aligning a tweet with research papers

D. Experiment: Calculation of Research Impact of Each Research Paper

Experimental Settings

In this section, we evaluate our ranking method described in Section III-C. In this experiment, we used 796 tweets and 237 research papers in 11 conferences, excluding the conferences for which no tweets were aligned with any awarded research papers. We conducted tests using our method and two baseline methods as follows.

- **Our method**: Count the number of tweets automatically aligned with the research papers using the DP + Sec method, and then sort research papers by the number of tweets.
- **Baseline**: Does not use the method for automatically classifying tweets for a research paper described in Section IV-B.

For the evaluation, we compared the output using our system with actual awards to research papers in each academic conference. This evaluation is based on Sidiropoulos's work [16]. As the evaluation measure, we used Mean Reciprocal Rank (MRR).

Results and Discussion

The experimental results are shown in Table IV. Our method obtained the higher MRR score of the two methods.

TABLE IV. EXPERIMENTAL RESULTS FOR CALCULATING RESEARCH IMPACTS OF EACH RESEARCH PAPER

	MRR
Our method	**0.236**
Baseline	0.183

This result indicates that our ranking method is useful for detecting research papers having higher research impacts.

We investigated the ranks of actually awarded research papers by our system in each academic conference. Examples of the ranking results of awarded research papers are shown in Table V. The numbers in parentheses following the text in the left column show the number of research papers in the conference. In the other columns, the numbers in parentheses show the actually awarded research papers for each award section. Rank-*n* shows the rank at which the awarded research paper appeared in our output list. If many of the awarded research papers appear at the top at the ranking list of the research papers in the conference, our system can be useful to find research papers having high research impact scores.

From Table V, our system could find several awarded research papers in Rank-1 such as #nlp2012 and #DEIM2012. In particular, our system shows the best performance in #nlp2012. This result shows that if we seek research papers having high research impact scores among the 29 research papers in #nlp2012, we may look at the top five in the ranking list provided by our system. However, many awarded research papers were not contained in the overall ranking results. For example, there are 13 awarded research papers in #jsai2010; however, our system found only one research paper (Rank-12) because no participants posted valuable tweets for the remaining 12. In our future work, we will therefore consider new metrics that find these research papers that our system could not find.

TABLE V. EXPERIMENTAL RESULTS FOR RANKING VALUABLE RESEARCH PAPERS

	Rank of awarded research paper		
	Best Paper Award	**Paper Award**	**Young Paper Award**
#jsai2010 (17)		Rank-12 (1/13)	
#jsai2012 (51)		Rank-22 (1/9)	
#rrds3 (7)	Rank-5 (1/1)	Rank-5 (1/1)	
#nlp2012 (29)	Rank-5 (1/1)	(0/4)	Rank-1, 5 (2/5)
(1) #sigwi2 (11)		Rank-9 (1/1)	(0/1)
(2) #sigwi2 (16)		Rank-11 (1/1)	Rank-14 (1/1)
(3) #sigwi2 (8)		Rank-6 (1/1)	(0/1)
(4) #sigwi2 (11)		Rank-1 (1/1)	(0/1)
#DEIM2012(63)	(0/1)	Rank-22 (1/1)	Rank-1, 30, 46 (3/15)
#DEIM2013(15)	(0/1)	(0/1)	Rank-3, 5, 12 (3/30)
#DEIM2014 (9)	(0/2)	Rank-1 (1/4)	(0/51)

V. CONCLUSION

In this paper, we proposed methods for automatically evaluating research impact using Twitter. We defined valuable and valueless tweets, and we proposed an approach based on the similarity between a tweet and a research paper. We also calculated research impacts of each research paper by the number of aligned tweets. To investigate the effectiveness of our methods, we conducted some experiments using tweets posted during several Japanese conferences and the research papers presented in them. From these results, we confirmed the effectiveness of our methods.

REFERENCES

[1] E. Garfield, "Citation Indexes for Science: A New Dimension in Documentation Thought the Association of Ideas", *Science*, pp. 108-111, 1955.

[2] J.E. Hirsch, "An Index to Quantify an Individual's Scientific Research Output", *Proc. the National Academy of Sciences of the United States of America*. Vol. 102, No. 46, pp. 16569-16572, 2005.

[3] D. Yogatama, B. Heilman, B. O'Connor, and C. Dyer, "Predicting a Scientific Community's Response to an Article", *Proc. the Conference on Empirical Methods in Natural Language Processing*, pp. 594-604, 2011.

[4] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation Count Prediction: Learning to Estimate Future Citations for Literature", *Proc. the 20th ACM International Conference on Information and Knowledge Management*, pp. 1247-1252, 2011.

[5] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee, "Towards a Stratified Learning Approach to Predict Future Citation Counts", *Proc. the 20th ACM International Conference on Information and Knowledge Management*, pp. 1247-1252, 2014.

[6] F. Davletov, A.S. Aydin, and A. Cakmak, "High Impact Academic Paper Prediction Using Temporal and Topological Features", *Proc. the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 491-498, 2014.

[7] L. Vaughan and D. Shaw, "Web Citation Data for Impact Assessment: A Comparison of Four Science Disciplines", *Journal of the American Society for Information Science and Technology*, pp. 1075-1087, 2005.

[8] H. Sayyadi and L. Getoor, "FutureRank: Ranking Scientific Articles by Predicting their Future PageRank", *Proc. the 9th SIAM International Conference on Data Mining*, pp. 533-544, 2009.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", *Technical report, Stanford Digital Library Technologies Project*, 1998.

[10] G. Eysenbach, "Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact", *Journal of Medical Internet Research*. Vol. 13. No. 4, 2011.

[11] K. Weller, E. Dröge, and C. Puschmann, "Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences", *Proc. the ESW2011 Workshop on Making Sense of Microposts*, pp. 1-12, 2011.

[12] J. Priem, D. Taraborelli, P. Groth, and C. Neylon, "Altmetrics", http://altmetrics.org/manifesto

[13] N. Kaji and M. Kitsuregawa, "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents", *Proc. the Conference on Empirical Methods in Natural Language Processing*, pp. 1075-1083, 2007.

[14] K.Y. Su, M.W. Wu, and J.S. Chang, "A New Quantitative Quality Measure for Machine Translation Systems", *Proc. the 14th Conference on Computational Linguistics*, pp. 433-439, 1992.

[15] C.Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries". Proc. the ACL-04 Workshop "Text Summarization Branches Out", pp. 74-81, 2004.

[16] A. Sidiropoulos and Y. Manolopoulos, "A Citation-Based System to Assist Prize Awarding", *Proc. the ACM SIGMOD Record*, Vol. 34, No. 4, pp. 54-60, 2005.