Hierarchical Estimation Framework of Multi-Label Classifying: A Case of Tweets Classifying into Real Life Aspects

Shuhei Yamamoto
Graduate School of Library,
Information and Media Studies,
University of Tsukuba, Japan
yamahei@ce.slis.tsukuba.ac.jp

Tetsuji Satoh
Faculty of Library,
Information and Media Studies,
University of Tsukuba, Japan
satoh@ce.slis.tsukuba.ac.jp

Abstract

Many people share their daily events and opinions on Twitter. Some are beneficial and comment on several aspects of a user's real life, i.e., eating, traffic conditions, weather, and so on. Since some tweets indicate two or more aspects, multi-label classification is required. Typical methods are not performed on tweets because they consist of short and elided sentences. To conquer these problems, we are researching a hierarchical estimation framweork (HEF) to estimate several aspects of unknown tweets. HEF is composed of both unsupervised and supervised machine learnings. In the first phase, it extracts topics from a sea of tweets using latent dirichlet allocation (LDA). In the second phase, it calculates the relevance between topcis and aspects using a small set of labeled tweets to build associations among them. In this paper, we introduce the entropy feedback method in the second phase. We evaluate the Shannon entropy of each association between the aspects and topics and iteratively calculate the feedback coefficients by entropy to achieve optimal associations. Our sophisticated experimental evaluations with a large amount of actual tweets demonstrate the high efficiency of our multi-labeling method. Our entropy feedback method successfully increased higher F-measures in all aspects. Expecially in *Disaster* and *Traffic* aspects, precision greatly increased without decreasing recall.

Introduction

Information sharing services continue to rapidly spread. At the end of 2013, Twitter boasted 200 million active users worldwide every month (Twitter 2014). Twitter only permits users to post short sentences up to 140 characters. Users can easily post about their experiences and share their opinions on daily events. Twitter posts are often both useful and timely because they are typically based on current events. For example, tweets about traffic jams or traffic accidents are quite valuable for users who will through those places. Supermarket sales and bargain information are also helpful for neighborhood consumers. Such tweets, which are highly regional, up-to-date, and beneficial to others, we call **real life tweets** (Yamamoto and Satoh 2013). For presenting such tweets based on user contexts, we classify them into 14 aspects. The 14 aspects shown in **Table 1** are assumed to be users life aspects that refer to the Yahoo directory [1] and so on. For example, a tweet that mentioned a traffic accident is labeled with a *Traffic* aspect, and another that mentioned supermarket sales and bargain information is labeled with an *Expense* aspect.

Depending on the tweets, we might have to estimate several aspects per tweet. For example, such a tweet as "A heavy snowstorm caused traffic accident near the JFK airport" mentions a heavy snowstorm and a traffic accident. Its main topic is the traffic accident, but it also provides weather information. Therefore, we label it as both *Traffic* and *Weather*.

In our previous research, we proposed a hierarchical estimation method to estimate several aspects of unknown tweets (Yamamoto and Satoh 2014). It is composed of both unsupervised and supervised machine learning techniques. In the first phase, it extracts topics from a sea of tweets using latent Dirichlet allocation (LDA). In the second phase, it calculates the relevance between topics and aspects using a small set of labeled tweets to build associations among them and aspect scores for unknown tweets using the associations between topics and aspects based on the terms extracted from them. When the aspect scores exceed a threshold, the aspects are estimated for tweets.

In this paper, we propose a hierarchical estimation framework (HFF) for achieving a seamless conjunction between the first and second phases. We also introduce entropy feedback mechanisms in the second phase to overcome the problem of competitive associations among aspects. Based on these extensions, the associations between topics and aspects are refined and the estimation precisions are increased. We evaluate the Shannon entropy of each association between the aspects and topics and iteratively calculate the feedback coefficients by entropy to achieve optimal associations. The relevance among them is recalculated using feedback coefficients. In our experimental evaluations, we show an improved estimation performance compared to our previous method. Moreover, we compare HEF with typical multi-labeling methods such as L-LDA, SVM, and NBML. After that, we discuss our method and conclude our research and briefly describe future works in later sections.

[1] http://business.yahoo.com

Table 1: Aspects of real life

Aspect		typical terms
Appearance	(App.)	clothes, dress, wearing, fashion, uniforms, kimono, decoration, makeup, haircuts ...
Contact	(Con.)	appointments, meetings, invitations, family, friends, parties, drinking parties, get-togethers ...
Disasters	(Dis.)	flood, tornados, earthquakes, seismic ocean waves, power loss, hazards, secondary disasters ...
Eating	(Eat.)	cooking, dining out, eating, restaurants, recipes, ingredients ...
Events	(Eve.)	festivals, ceremonies, projects, schedules of events, conferences, special days, art shows ...
Expense	(Exp.)	shopping, orders, advertisements, discounts, bargains, markets, sales, purchases ...
Health	(Hea.)	colds, physical condition, aches and pains, hospital, health management method, medicine ...
Hobbies	(Hob.)	leisure-time, pastime, entertainment, hobbies, interest, games, music, television, movies ...
Living	(Liv.)	home, lodgings, furniture, cleaning, doing laundry, living, apartment, accommodation ...
Locality	(Loc.)	sightseeing, regionally specific, local information ...
School	(Sch.)	study, class, examinations, education, research, homework, coursework, lectures cancellation ...
Traffic	(Tra.)	trains, buses, airplanes, timetables, traffic information, clogs, roads, traffic jams, accidents ...
Weather	(Wea.)	weather forecasts, temperature, humidity, hail, rain, thunder, sky, air, wind, pollen ...
Work	(Wor.)	job hunting, part-timer, coursework, opening a store, closing a business, job, employment ...

Related Works

Information extraction from Twitter

The study of information extraction from Twitter is flourishing. Sakaki et al. (Sakaki, Okazaki, and Matsuo 2010) assumed that Twitter users act as sensors that discover an event occurring in real time in the real world. Mathioudakis et al. (Mathioudakis and Koudas 2010) extracted burst keywords in automatically collected tweets and found trends that fluctuated in real time by creating groups using the co-occurrence of keywords. Zhao et al. (Zhao and Mei 2013) extracted tweets about information needs using a Support Vector Machine (SVM) to discover real world trends and events. Wang et al. (Wang et al. 2013) estimated user interests using posted tweets to discover effective users for tweet diffusion. In this paper, we estimate real life aspects of unknown tweets.

Topic model

Topic model studies widely use LDA (Blei, Ng, and Jordan 2003), which is a latent topic extracting method that was devised for a probability topic model. LDA supposes that a document is a mixture distribution of plural topics. Each topic is expressed by the probability distribution of the terms. Zhao et al. (Zhao et al. 2011) proposed a model called Twitter-LDA, based on the hypothesis that one tweet expresses one slice of a topic's content. They classified tweets by topics and extracted keywords to express their contents. Zhang et al. (Zhang et al. 2012) recommended bands to music lovers using LDA by calculating the degree of artist similarity based on generated topics. Users received recommendations about artists in whom they might be interested. Riedl et al. (Riedl and Biemann 2012) found the change-points of topics using LDA by calculating the similarity between sentences that express the vectors of topic frequency. In this paper, we build associations between aspects and topics generated by LDA.

Multi-label classification

Multi-label classification studies are widely known methods based on SVM, naive Bayes classifiers, and LDA. SVM, which is one identification method that performs supervised learning, has high generalizing capability and classification performance (Cortes and Vapnik 1995). Chang et al. (Chang and Lin 2011) developed a SVM library called LIBSVM, which achieves multi-label classification by building models by combining several labels.

A naive Bayes classifier assumes that the term occurrence in a document is independent, and label probabilities are calculated from these terms using Bayes rules. It estimates labels with the highest probability for a document (Domingos and Pazzani 1997). Wei et al. (Wei et al. 2011) proposed multi-label classification based on naive Bayes classifiers and estimated several labels with the probability that exceeds the average score calculated by all the label probabilities.

Ramage et al. (Ramage et al. 2009) suggested a model called Labeled LDA (L-LDA) that expanded LDA to supervised learning. To extract latent topics, it assumes the labels to be the contents of documents. L-LDA can extract a one-to-one correspondence between LDA's latent topics and document labels.

These methods show high estimation performance of such long documents as blogs and newspapers using sufficient training data. However, tweets consist of fewer terms because their length averages 45 characters (Mizunuma et al. 2014). Moreover, as training data, fresh tweets are preferred because they are easily influenced by the real world. In these conditions, typical multi-label classification methods fail to produce adequate performance to estimate several aspects of unknown tweets (Yamamoto and Satoh 2014).

Hierarchical Estimation Framework

Overview of HEF

In our previous research (Yamamoto and Satoh 2014; 2013), we estimated several aspects of a tweet by implementing the

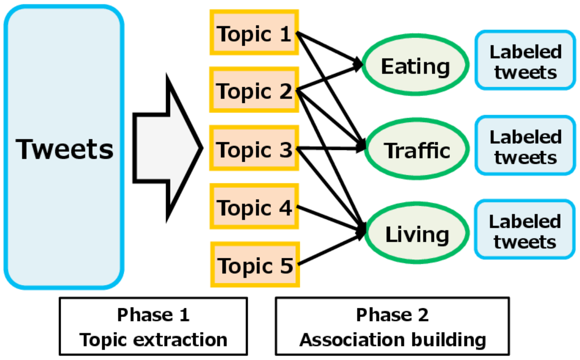


Figure 1: Hierarchical estimation framework

hierarchical estimation method as a fundamental part of the hierarchical estimation framework (HEF) (**Fig. 1**).

In the first phase of HEF, a large numbers of topics are extracted from a sea of tweets using LDA. In its second phase, associations between topics and aspects are constructed using a small set of labeled tweets. We calculated the aspect scores for unknown tweets using the associations based on the terms extracted from them. Appropriate aspects are used to label unknown tweets by certain thresholds.

Typical supervised machine learning methods directly calculate the term likelihood from labeled training data. The terms in unknown tweets, which do not appear in the training data, can't play a role in the estimation of previous methods. In contrast, HEF is composed of a triple hierarchy: Tweet-Topic-Aspect. The terms in a tweet are expanded using co-occurrence terms in appropriate topics. Thus, we believe that this feature is crucial for estimating with small sets and short sentences of labeled data: i.e., tweets.

In this paper, we introduce entropy feedback mechanisms in the second phase of HEF to conquer the problem of competitive associations among particular aspects. According to these extensions, associations between topics and aspects are refined and estimation precision will increase.

Association building

We build associations between many topics and fewer aspects. For building associations, we prepared a small set of labeled tweets. A set of extracted terms from tweets is W. We extracted nouns, verbs, and adjectives using a Japanese morphological analyzer called MeCab (Kudo 2005).

Relevance $R(a,t)$ between topics t and aspects a is calculated as follows:

$$R(a,t) = \sum_{w \in W} p(a,w)^{\alpha} * p(t,w)^{\beta}, \qquad (1)$$

where $p(t,w)$ denotes the occurrence probability of term w in topic t. $p(a,w)$ denotes the occurrence probability of term w in aspect a calculated by a small set of labeled tweets. Note this equation only calculates the relevance between topics and aspects using the occurrence probability. α and β, which are feedback coefficients to control the extent of occurrence probability, are calculated in next section.

To fold value 0 into 1, we normalize $R(a,t)$ in each aspect and topic. The normalized relevances in each aspect $\hat{Ra}(a,t)$ and normalized relevance in each topic $\hat{Rt}(a,t)$ are shown as follows:

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum\limits_{x \in T} R(a,x)}, \hat{Rt}(a,t) = \frac{R(a,t)}{\sum\limits_{x \in A} R(x,t)}, \qquad (2)$$

where T denotes all topics extracted using LDA. A is all the aspects. $\hat{Ra}(a,t)$ is a representation feature where aspect a is supported from topic t. $\hat{Rt}(a,t)$ is a representation feature where topic t supports aspect a.

We make associations between topics and aspects. Here, depending on the aspects, note that the associations with topics are different. For example, the Eating aspect may be supported by fewer topics with high probabilities, and the Living aspect may be supported by many topics with mid-level probabilities (**Fig. 1**). We must construct various associations of each aspect because the optimal topic set is different for each aspect.

Therefore, we make an association between topics and aspects when $\hat{Ra}(a,t)$ exceeds a calculated threshold in each aspect a. Topic set T_a of aspect a is shown as follows:

$$T_a = \{t | \hat{Ra}(a,t) > \max_{t \in T} \hat{Ra}(a,t) - \sigma(\hat{Ra}(a,T)) * d\}, \quad (3)$$

where $\sigma(\hat{Ra}(a,T))$ denotes the standard deviation in $\hat{Ra}(a,t)$ for all topics.

According to increase the parameter d, aspects are associated to more topics. The optimal value of d is caused when associations between topics and aspects achieve the maximum estimation performance.

Entropy feedback

In our previous research (Yamamoto and Satoh 2013), we clarified that topics are competitive among particular aspects. The Disaster, Event, Locality, and Traffic aspects were associated with similar topics, which had such regional names as "Kyoto" and "Shijo". Tweets indicating these aspects probably appear in the regional names in a sentence because real life tweets mention the real world. This problem caused misguided estimations and lowered estimation precision.

LDA primarily provides high occurrence probability for high frequency terms in the dataset. Regional location terms have higher occurrence probability because they often appear in tweets. Therefore, to accurately calculate the relevance between topics and aspects, HEF has two kinds of parameters, α and β (Eq. (1)). In this paper, we propose a feedback method using Shannon entropy (Shannon 1951) to determine these parameters.

Entropy can evaluate the untidiness of probability distribution. $\hat{Ra}(a,\cdot)$ and $\hat{Rt}(\cdot,t)$ express the probability distribution in each aspect a and topic t. The entropies of both $H(a)$ and $H(t)$ are defined as follows:

$$
\begin{aligned}
H(a) &= -\sum_{t \in T} \hat{R}a(a,t) * \log_2 \hat{R}a(a,t), \\
H(t) &= -\sum_{a \in A} \hat{R}t(a,t) * \log_2 \hat{R}t(a,t). \qquad (4)
\end{aligned}
$$

Here, we must consider the association balance from some topics to an aspect. For example, as mentioned above, if such special terms as location names have high occurrence probability, the relevance is greatly high and entropy is low. Such association creates an unbalance for all the aspects. Hence, to control the occurrence probability of the terms, we calculate the feedback coefficients of both α and β on the basis of minimum entropy. α and β are calculated as follows:

$$
\alpha = \frac{1}{|A|} \sum_{a \in A} \frac{\min_{x \in A} H(x)}{H(a)}, \beta = \frac{1}{|T|} \sum_{t \in T} \frac{\min_{x \in T} H(x)}{H(t)}, \qquad (5)
$$

where $|A|$ and $|T|$ denote the number of aspects and topics.

If the entropy difference of all the aspects and topics is increased, α and β are decreased. When both feedback coefficients are introduced to Eq. (1) within 1.0, the difference of the occurrence probability in the topics or the aspects is reduced; α and β lower the effectivity of the terms with especially high occurrence probability, such as place names. As a result, the entropy difference of every aspect and topic decrease, and the association balance of every aspect is preserved. Suitable associations between aspects and topics are built when α and β converge.

HEF is iteratively calculated in the order of Eqs. (1), (2), (4), and (5). When α and β sufficiently converge compared to previous iteration values, HEF builds associations between topics and aspects by Eq. (3).

Estimation aspects for tweets

To estimate the aspects of unknown tweets, we use the associations between topics and aspects. The estimation flow using the associations is shown in **Fig. 2**. First, nouns, verbs, and adjectives are extracted from tweets. Second, the occurrence probabilities of all the terms are calculated for each topic. Then, the aspect score is calculated based on the tweet's probabilities and associations. Aspect scores $S(tw, a)$ between tweets tw and aspects a are calculated as follows:

$$
S(tw, a) = \sum_{t \in T_a} \sum_{w \in W_{tw}} p(t, w)^{\beta} * \hat{R}a(a,t) * \hat{R}t(a,t), \qquad (6)
$$

where W_{tw} denotes a set of terms extracted from unknown tweet tw and $p(t, w)$ denotes the occurrence probability of terms w in topic t. β denotes the feedback coefficient calculated by Eq. (5).

$\hat{R}a(a,t)$ gives high relevance to important topics for aspects. However, several aspects might strongly associate with the same topics. For example, topics in which verbs have a high rank of occurrence probability are given high relevance from many aspects because verbs often appear in many aspects. We believe that these topics decrease the estimation precision of aspects. $\hat{R}t(a,t)$ also gives high relevance to the characteristic topics of aspects, and low relevance to topics that share several aspects. Here, we must consider the properties of real life aspects with examples. For example, flood and heavy rain often appear in the same sentence because floods are generally caused by heavy rain; they are aggregated in the same topic by LDA. From **Table 1**, because flood and heavy rain are respectively included in Disaster and Weather aspects, both should share flood and heavy rain topics. However, $\hat{R}t(a,t)$ gives low relevance to Disaster and Weather aspects.

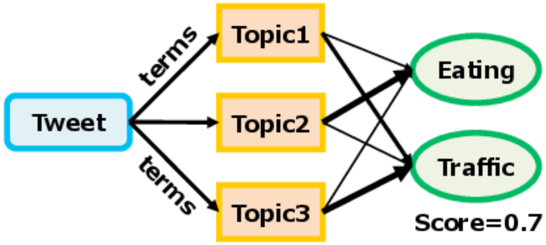


Figure 2: Aspect estimation method

To consider the relevance of both $\hat{R}a(a,t)$ and $\hat{R}t(a,t)$, we multiply both the relevances of the score calculation with Eq. (6).

Aspects with high scores should be estimated for tweets. We estimate the top K aspects that are flexibly decided. In HEF, each aspect a score $S(tw, a)$ is normalized using score average $\mu(S(tw, A))$ and standard deviation $\sigma(S(tw, A))$. If the normalized aspect score exceeds each aspect's threshold $r(a)$, aspects are more likely to be estimated for the tweet. Some aspects A_{tw} for unknown tweet tw are estimated as follows:

$$
A_{tw} = \left\{ a \,\middle|\, \frac{S(tw, a) - \mu(S(tw, A))}{\sigma(S(tw, A))} > r(a) \right\}. \qquad (7)
$$

Depending on the aspects, the estimation probabilities of the labels are intrinsically different. HEF decides threshold $r(a)$ in each aspect a from the number of labels $L(a)$ in the training data. Each aspect threshold $r(a)$ is calculated as follows:

$$
r(a) = \frac{\mu(L(A)) - L(a)}{\sigma(L(A))}, \qquad (8)
$$

where $\mu(L(A))$ and $\sigma(L(A))$ denote both labels average and standard deviations of labels. This equation subtracts the number of each labeling aspect $L(a)$ from average value $\mu(L(A))$; the threshold is high when the number of labelings is less, and it is low when the number of labelings is great.

Optimal number of topics

LDA needs the number of topics as a parameter, which is important for our method because associations between topics and aspects are based on relevance. If the number of topics changes, the number associated with the aspects also changes.

To select the best number of topics in LDA, we used the JS Divergence (Murphy 2012) between each aspect and applied it to calculate the similarity between one aspect and others. When the JS Divergence is high, the probability distribution among aspects is much different. When it is 0, the probability distribution is identical. In this case, the maximum value of the JS Divergence sum indicates the optimal aspect set. Probability distributions use the $\hat{Ra}(a, t)$ of the aspects and the topics matrix. JS Divergence sum JS_{sum} is calculated as follows:

$$\begin{aligned} JS_{sum} &= \sum_{(\forall p, \forall q) \in A} D_{JS}(\hat{Ra}(p, \cdot), \hat{Ra}(q, \cdot)), \qquad (9) \\ D_{JS}(x, y) &= \frac{1}{2}\left(\sum_{t \in T} x(t) \log \frac{x(t)}{z(t)} + \sum_{t \in T} y(t) \log \frac{y(t)}{z(t)}\right), \end{aligned}$$

where $z(t)$ denotes the average of $x(t)$ and $y(t)$.

Experimental Evaluations

To clarify the effectiveness of our HEF which introduced feedback entropy method, we evaluated the precision, recall, and the F-measure values of the estimated aspects. As baseline methods, we used L-LDA, SVM, and NBML. By analyzing the associations between topics and aspects, we clarified the aspects for which the entropy feedback method was effective.

Dataset and parameter settings

Collecting many regional tweets: Our method requires many tweet datasets for generating topics using LDA. We collected 2,390,553 tweets posted from April 15, 2012 to August 14, 2012 using the Search API [2] on Twitter, each of which has "Kyoto" as the Japanese location information.

Real life tweets: To construct associations between the extracted topics and aspects, we prepared a small set of 1,500 labeled tweets, each of which has "Kyoto" as the Japanese location information. We used three examinees: examinee E1 is the first author, and E2 and E3 are university students living in Tsukuba City. During the labeling process, the examinees freely consulted **Table 1** and viewed the example tweets in each aspect and why they were classified as such. They selected the most suitable aspect for each tweet as the first aspect and the next two most suitable aspects as the second and third aspects. If no suitable aspect remained, they selected "other" to identify it as a non-real life tweet. Aspects that do not correspond to any candidate are listed fourth.

We evaluated the κ coefficients among the first candidates of the examinees (Cohen 1960). When the κ coefficient is high, the classification agreement rate among the examinees is also high. The κ coefficient for examinees E1 and E2 was 0.687; it was 0.595 for examinees E1 and E3 and 0.576 for examinees E2 and E3. The average was 0.619, which is a *substantial* match rate.

To appropriately give aspects to each tweet, we used the results from the labeling of all three examinees. Correct aspects AC_{tw} of each tweet tw are shown as follows:

$$\begin{aligned} AC_{tw} &= \{a | Uscore(tw, a) \leq 10\}, \\ Uscore(tw, a) &= \sum_{u \in U} candidate(tw, a, u), \qquad (10) \end{aligned}$$

where U denotes all the examinees. $candidate(tw, a, u)$ is a candidate number: the 1st, 2nd, 3rd, and 4th rankings of aspects a labeled by examinee u for tweet tw. Hence, maximum $Uscore(tw, a)$ is 12 with three examinees when all $candidate(tw, a, u) = 4$. Minimum $Uscore(tw, a)$ is three when all $candidate(tw, a, u) = 1$.

For this determination, the number of labeling aspects of 1,500 tweets is shown in **Table 2**. The number of labels in the Appearance aspect is 181. The minimum number of labels is 86 in the Disaster aspect. The number of all labels for 1,500 tweets is 5,092, and the per tweet average of the labels was 3.39.

Parameter settings: LDA requires hyperparameters. Based on related works (Griffiths and Steyvers 2004), we set α to $\frac{50}{|T|}$ and β to 0.1. $|T|$ denotes the number of topics, chosen based on JS_{sum} from among 50, 100, 200, 500, and 1,000 topics in the *Number of topics* section. The iterative calculation count in LDA is 100 times in every case.

Baseline methods

We prepared such typical multi-label classification methods as L-LDA (Ramage et al. 2009), LIBSVM (Chang and Lin 2011), and NBML (Wei et al. 2011) for evaluating HEF's effectiveness with the entropy optimizations.

LIBSVM requires some parameters. We chose a linear kernel and set parameter C to 1.0, indicated by a grid search in the LIBSVM tools(Hsu, Chang, and Lin 2010). The features for all the methods are nouns, verbs, and adjectives, which were obtained by morphological analysis.

L-LDA has to set the hyperparameters of both α and β, like in LDA. We experimentally set α to 0.1 and β to 0.1, and the iterative calculation count in L-LDA was 100.

Experimental results

Number of topics: We evaluated JS_{sum} to tune the number of topics. The list of JS_{sum} that varies the number of topics is shown in **Fig. 3**. The maximum value appears in 500 topics. We concurrently evaluated the precision, the recall, and the F-measure in each topic. The maximum precision and recall were achieved in 200 and 1,000 topics, and the maximum F-measure was achieved in 500 topics. Therefore, we used 500 as the optimal number of topics for HEF. The decision method of the optimal number of topics by the JS_{sum} value is generally effective for HEF because stable evaluation values were achieved in about 500 topics.

[2] https://dev.twitter.com/docs/api/1/get/search

Table 2: Number of correctly labeled aspects

Aspect	App.	Con.	Dis.	Eat.	Eve.	Exp.	Hea.	Hob.	Liv.	Loc.	Sch.	Tra.	Wea.	Wor.	Other	Total
Label	181	379	86	287	311	435	177	348	213	432	195	169	226	262	1,391	5,092

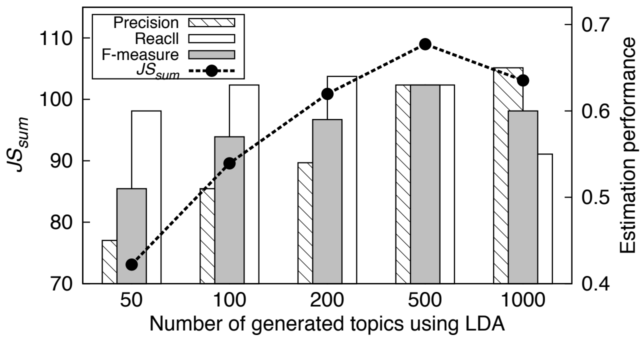


Figure 3: JS_{sum} value of each number of topics

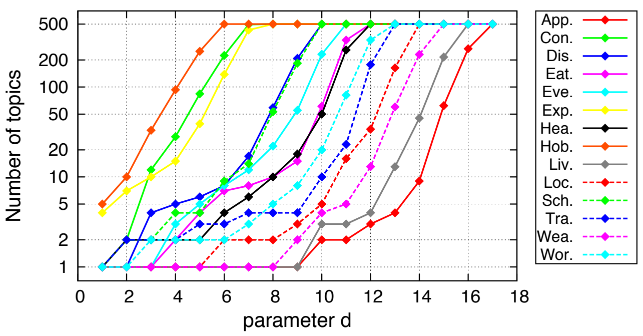


Figure 5: Connectivity among topics and aspects

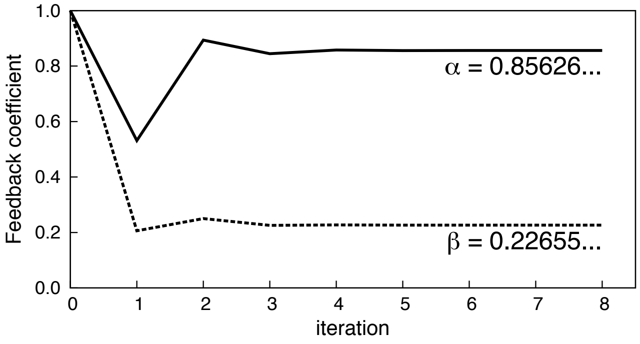


Figure 4: Converging state of feedback coefficients

Feedback coefficients: The feedback coefficients of both α and β vary, as shown in **Fig. 4**. The convergence condition was set within a difference of 0.00001 compared to previous iteration values. The starting value was set to 1.0. Both converged at eight iterations. From this result, α and β became 0.85626 and 0.22655.

Connections from topics to each aspect: To analyze the association between topics and aspects, we evaluated the number of connections from the topics to each aspect. The number of topics connecting each aspect varying to parameter d is shown in **Fig. 5**. In all aspects, the number of topics increased based on d. The Appearance aspect is most closely connected to one topic, $d \leq 11$. The Hobby aspect connects to much topics with fewer value of d, and it completely connects to all the topics at $d = 6$. When d exceeds 18, the associations between topics and aspects become a complete bipartite graph.

Transitions of precision, recall, and F-measure values: The precision, recall, and F-measure values are shown in **Fig. 6** for the aspects of Disaster, School, and Traffic. The horizontal axis is parameter d, which decides the association between topics and aspects.

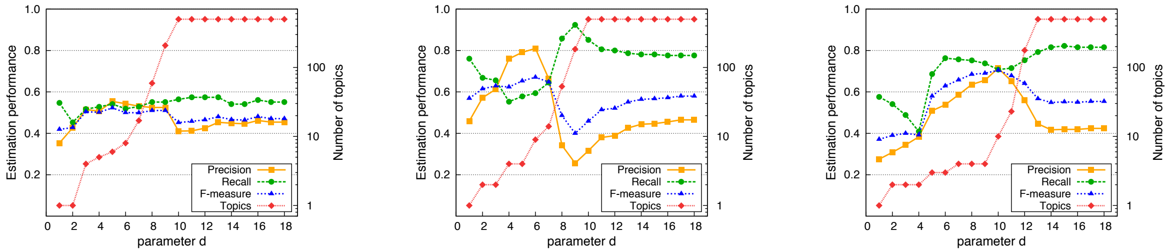
In the Disaster aspect, recall slowly increased based on d. In contrast, precision decreased based on d. The maximum F-measure was achieved at $d = 5$.

In the School aspect, precision rapidly increased until $d \leq 6$ and then quickly decreased until $d \leq 9$. Recall decreased until $d \leq 4$ and then increased until $d \leq 9$. The maximum F-measure was achieved at $d = 6$.

In the Traffic aspect, the precision, recall, and F-measure values increased until $4 \leq d \leq 6$. There are three topics at $d = 6$. Precision increased until $d \leq 10$ and then decreased until $d \leq 14$. The maximum F-measure was achieved at $d = 10$.

The evaluation values change even after they are connected to all the topics in these aspects. The associations of other aspects change based on an increased d until $d \leq 18$, and the aspect scores also change. We show the optimal d of each aspect in the next section.

Estimation performance of each method: The precision, recall, and F-measure values of each method are shown in **Table 3**. All of the methods were evaluated using 10-fold cross validations. In each evaluation, 1,350 tweets were used for model training, and the remaining 150 tweets were used to evaluate the precision, recall, and F-measure values. We also calculated their macro averages. The highest value in each row is shown in bold. The HEF columns show our method, where associations were built using entropy feedback explained in the *Feedback coefficients* section. The HEF0 columns show the 0 iteration cases of entropy feedback, and both α and β are 1.0. Optimal values of d when achieved the highest F-measure are shown in the far right d column in **Table 3**. Optimal d of Appearance is 17 when precision and recall are 0.74 and 0.53. In the Disaster and



(a) Disaster (b) School (c) Traffic

Figure 6: Precision, Recall, and F-measure evaluated by varying parameter d

Table 4: Number of labelings by each method

Labels	HEF	L-LDA	SVM	NBML	Examinees
1	0	101	0	165	1
2	0	154	137	531	111
3	22	259	1,250	442	820
4	389	369	80	243	442
5	574	307	33	90	115
6	382	182	0	23	11
7	118	80	0	6	0
8	15	37	0	0	0
9	0	9	0	0	0
10	0	2	0	0	0
Average	5.15	4.16	3.00	2.75	3.39

Traffic aspects, HEF's precisions greatly increased without decreasing recall more than HEF0's. Disaster's F-measure by HEF surpassed 0.25 points ($= 0.54 - 0.29$) compared with HEF0. HEF's average F-measure showed the highest value in all the methods.

The number of labels, each of which was estimated as an aspect of the tweets by all methods and the examinees, is shown in **Table 4**. In the examinees and SVM, there are three labeling modes. The maximum and minimum numbers of labeling modes in every method are found in the HEF and NBML values.

Topics associated with each aspect: Next we examined the topics connected to each aspect. The associations built by HEF0 are shown in **Table 5**. This table shows the topic ids of top four that are strongly connected to each aspect. Three or more times appearing topics in every aspect are marked in bold. For example, the Appearance aspect is associated to topic 119 with highest relevance $\hat{Ra}(a,t)$. Topic 125 associates to the Disaster, Event, Locality, and Traffic with the highest (1st rank) relevances and Weather with the second highest (2nd rank) relevance. Moreover, topics 125, 299, and 469 appear together in the Disaster, Event, and Locality aspects. Topic 60 appears in the aspects of Disaster, Locality, and Traffic.

Similarly, the associations built by HEF are shown in **Table 5**. Note that these topic ids are same as HEF0's topic ids. The associations built by HEF are quite different from those by HEF0. For example, topic 125 appears only in the Locality aspect with 4th rank unlike in HEF0's associations. The aspects of Disaster and Event are associated topic 178 and 345 with 1st rank. Topic 60 connects to both aspects of Locality and Traffic with 1st rank.

Estimation precision using a small bit of labeled data: In all the methods, we evaluated the estimation performance using less training data. We split the datasets into 10 subsets, and only one subset is circularly selected as a test dataset. From the remaining nine subsets, we randomly extracted 1 set (150 tweets), 3 sets (450), 5 sets (750), and 7 sets (1050) as the training data. We calculated the average evaluation value by repeating ten times changing the test data. Each evaluation value is shown in **Fig. 7**. We chose optimal d as the HEF parameters, as in the *Estimation performance of each method* section. The optimal number of the topics in all the training data was 500, depending on JS_{sum}.

HEF's precision is lower than NBML's with training dataset 9 (1350). However, based on the decreasing training data, the precision difference of both methods was small. The precision of our method did not fall even when the amount of training data decreased. In recall, the precision of L-LDA and SVM rapidly fell with less training data; however, HEF and NBML showed almost no drop. In the F-measures, HEF achieved the high score until training dataset 3, and it is usually the maximum F-measure in all the methods. The F-measure of SVM rapidly dropped with less training data.

Discussions

Effectiveness of feedback entropy

According to **Table 3**, HEF's F-measure increased more than HEF0's F-measure in many aspects, especially for Disaster, Traffic, and Weather F-measure values. From **Table 5**, Disaster is strongly associated to Topics 125, 299, 460, and 60. Part or all of them were also associated with Event, Locality, Traffic, and Weather aspects. The characteristic words in HEF0 are shown in **Table 6**. Topic 125 has "aquarium" and "tower" that denote the names of structures, and "Yamashina" and "Sakyo" denote place names. Topic 125 is related to the names of geographic elements around Kyoto. Topic 299 is related to tourism/tourists in Kyoto because it includes "sightseeing" and "travel." Topic 469 is related to living in Kyoto because "welfare" and "nursing" are found in it. In these topics, "Kyoto" exists at the top priority of the characteristic words. Therefore, these topics describe the

Table 3: Precision, Recall, and F-measure of each method

	Precision					Recall					F-measure					
Aspect	HEF0	HEF	L-LDA	SVM	NBML	HEF0	HEF	L-LDA	SVM	NBML	HEF0	HEF	L-LDA	SVM	NBML	d
Appearance	0.61	0.74	0.43	0.64	**0.82**	0.54	0.53	**0.69**	0.28	0.37	0.57	**0.62**	0.52	0.38	0.51	17
Contact	0.40	0.34	0.43	0.41	**0.53**	0.55	**0.71**	0.68	0.35	0.54	0.46	0.46	**0.53**	0.37	**0.53**	3
Disaster	0.21	0.55	0.67	0.44	**0.76**	0.49	**0.54**	0.49	0.44	0.21	0.29	**0.54**	**0.54**	0.44	0.33	5
Eating	0.66	**0.75**	0.41	0.51	0.73	0.74	**0.77**	**0.77**	0.64	0.51	0.70	**0.76**	0.53	0.57	0.60	8
Event	0.41	0.39	0.47	**0.56**	**0.56**	0.55	**0.62**	0.51	0.20	0.45	0.47	0.48	0.48	0.29	**0.49**	10
Expense	0.39	0.47	**0.64**	0.43	0.52	**0.65**	0.57	0.40	0.45	0.46	0.49	**0.51**	0.49	0.43	0.49	13
Health	0.31	0.62	0.43	0.48	**0.76**	**0.56**	0.38	0.55	0.28	0.38	0.40	0.46	0.48	0.35	**0.50**	10
Hobby	0.32	0.32	0.44	0.43	**0.57**	0.84	**0.87**	0.62	0.54	0.44	0.46	0.47	**0.51**	0.48	0.49	17
Living	0.34	0.63	0.38	0.64	**0.71**	**0.74**	0.50	0.62	0.34	0.41	0.46	**0.55**	0.46	0.44	0.51	16
Locality	**0.65**	**0.65**	0.62	0.62	0.62	0.66	**0.73**	**0.73**	0.54	0.65	0.65	**0.69**	0.67	0.57	0.63	10
School	0.57	0.81	0.37	**0.88**	0.81	0.59	0.59	**0.81**	0.36	0.52	0.58	**0.67**	0.51	0.49	0.63	6
Traffic	0.54	0.72	0.33	0.71	**0.82**	0.68	0.71	**0.82**	0.44	0.50	0.60	**0.71**	0.47	0.54	0.62	10
Weather	0.28	**0.89**	0.25	0.47	0.81	0.81	0.50	**0.84**	0.63	0.58	0.41	0.64	0.38	0.53	**0.67**	5
Working	0.38	**0.69**	0.64	0.52	0.56	**0.64**	0.36	0.50	0.19	0.35	0.47	0.47	**0.55**	0.28	0.43	12
Other	0.93	0.93	**0.94**	0.93	0.93	**0.99**	**0.99**	0.51	**0.99**	0.93	**0.96**	**0.96**	0.66	**0.96**	0.93	1
Average	0.47	0.63	0.50	0.58	**0.70**	**0.67**	0.63	0.63	0.44	0.49	0.55	**0.63**	0.52	0.47	0.56	

Table 5: High relevance $\hat{Ra}$ topics in each aspect built by HEF0 and HEF.

(a) HEF0

Aspect	1st rank	2nd rank	3rd rank	4th rank
Appearance	#119	#368	#458	#164
Contact	# 49	# 9	#157	**#490**
Disaster	**#125**	**#299**	**#469**	**# 60**
Eating	#207	#197	#484	#352
Event	**#125**	#345	**#299**	**#469**
Expense	#437	#454	# 11	#223
Health	#237	#359	#479	#393
Hobby	#221	#332	#311	#497
Living	#290	#275	#301	# 11
Locality	**#125**	**#299**	**# 60**	**#469**
School	# 3	**#490**	#443	#275
Traffic	**#125**	**# 60**	**#299**	#201
Weather	#451	**#125**	**#490**	#230
Working	#334	#253	# 21	#463
Other	#237	#359	#479	**#490**

(b) HEF

Aspect	1st rank	2nd rank	3rd rank	4th rank
Appearance	#119	#474	#240	#454
Contact	# 49	#429	#157	#466
Disaster	#178	#380	**#469**	#277
Eating	#341	#484	#352	#207
Event	#345	#314	#190	#307
Expense	#437	# 35	#454	#419
Health	#393	# 22	#348	#193
Hobby	# 75	#412	#273	#430
Living	#290	#133	#230	#301
Locality	**# 60**	#314	**#299**	**#125**
School	# 3	#111	#118	#418
Traffic	**# 60**	#201	#149	# 42
Weather	# 23	#451	**#490**	#178
Working	#321	#436	#253	#334
Other	#281	#330	#304	# 21

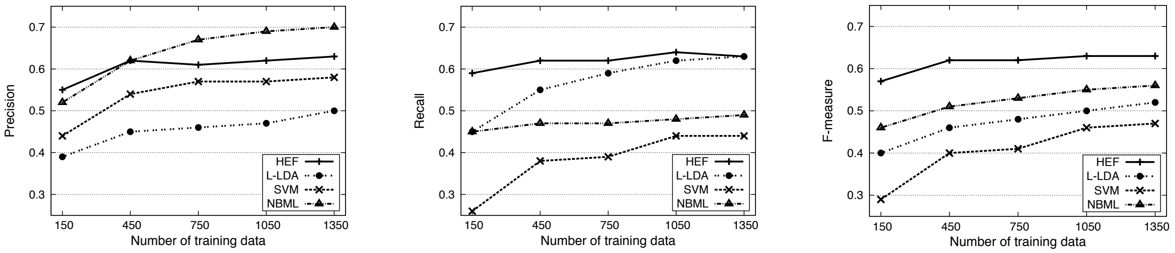
Kyoto district and are connected to the Locality aspect; however, they are also connected to other aspects, such as Disaster, Event, and Traffic. To explain these diverse connections, Disaster or Event tweets frequently include such place words. For example, earthquake tweets usually describe not only the earthquake itself but also its center, which is obviously a geographic name. In topics about districts, geographic names have very high occurrence probability, as shown in **Table 6**. For these reasons, similar sets of topics are connected to the Disaster, Event, and Traffic aspects.

On the other hand, for the associations built by HEF, almost none of these topics appeared in all of the aspects (**Table 5**). The most strongly associated topics with Disaster, Locality, Traffic, and Weather are Topics 178, 60, and 23, respectively. Their characteristic words are shown in **Table 6**. Topic 178 has "typhoon" and "storm," which denote natural disasters, and the tweet was associated with a Weather aspect in $\hat{Ra}$ rank 4. Topic 60 has such place names as "Kyoto" and "Kawaramachi." It also has "subway" and "city bus," which are usually used for the Traffic aspect. For these reasons, the Locality and Traffic aspects share Topic 60. Topic 23 has "sunny" and "forecast," which are usually used for the Weather aspect. These relationships among characteristic words and real life aspects are also shown in **Table 1**. From these results, higher F-measures in the aspects of Disaster, Locality, Traffic, and Weather are achieved by strongly connected topics: Topics 178, 60, and 23.

Estimation performance of each method

From **Table 3**, HEF's average precision (0.63) is lower than NBML's. But HEF's average recall (0.63) and its average F-measure (0.63) are higher than NBML's. In **Table 4**, the number of labelings by NBML is the lowest in all the methods and fewer than the labels of the examinees. NBML's precision rose but its recall fell. When we compare the average recalls of HEF and L-LDA, we see that HEF's recall is the same as L-LDA's. From **Table 4**, HEF estimates more labels than L-LDA. However, its precision and its F-measure are higher than L-LDA's. Since HEF estimated more correct aspects than L-LDA, our method accurately calculated the aspect scores of tweets.

From **Fig. 7**, our method shows the results where the descent of the precision is small. The recalls of HEF and L-LDA have almost no difference with training dataset 9. With less training data, L-LDA's recall rapidly dropped. However, HEF showed almost no drop. In every method except HEF, the recall values rapidly fell because the terms decreased based on less training data. On the other hand, in



(a) Precision (b) Recall (c) F-measure

Figure 7: Precision, Recall, and F-measure evaluated by varying amount of training data

Table 6: High occurrence probability terms in each topic

Topic	Characteristic words
Topic 125	Kyoto, newspapers, city centers, aquariums, towers, Yamashina, Sakyo, living, Fushimi
Topic 299	Kyoto, sightseeing, hotels, taxis, roaming, travel, school trips, lodging
Topic 469	Kyoto, citizens, institutions, environments, nursing, welfare, newspapers, medical
Topic 178 Dis. $\hat{Ra}$ rank 1	typhoons, Kyoto, alarm, heavy rain, storms, influence, precautions, floods
Topic 60 Loc. $\hat{Ra}$ rank 1	Kyoto, traffic, Kawaramachi, Shijyo, Torimaru, subways, guides, city buses
Topic 23 Wea. $\hat{Ra}$ rank 1	weather, sunny, forecast, Kyoto, rainy season, clouds, temperature

our method, topics are associated to aspects. Therefore, the terms don't decrease even if the number of training data decreased. For these reasons, the HEF's recall almost didn't fall based on less training data. Hence, the HEF's F-measure is higher than all the other methods.

A sample tweet is shown in **Table 7**. The examinee aspect column shows the aspects labeled by the examinees, based on Eq. (10). The columns of the HEF, HEF0, and NBML aspects estimated the aspects by each method. **Table 7** is a completely matched example whose aspects were labeled by the examinees and estimated by HEF. It shows the effectivity and the characteristics of HEF estimation and mentions a restaurant's opening in "Takaragaike". The examinee aspects are Eating, Expense, and Locality, all of which coincide with the tweet's topics. NBML estimated Eating and Expense aspects but failed to estimate the Locality aspect because it was not trained by the likelihood between "Takaragaike" and Locality by the training data. HEF0 estimates many aspects: Eating, Expense, Locality, Disaster, Event, and Traffic. Obviously, this tweet does not mention Traffic, Disaster, or Event. HEF0 excessively estimated aspects because it built associations between many aspects and local topics, such as Topic 125. On the other hand, the aspects estimated by HEF match the examinees' aspects. HEF estimated Locality because it built associations between the Locality aspect and a topic including "Takaragaike". Since other aspects were not associated to the local topics, HEF accurately estimated the aspects.

Table 7: Complete estimated aspects for tweet by HEF

Answers	Loc., Exp., Eat.
HEF	Loc., Exp., Eat.
HEF0	Loc., Exp., Eat., Eve., Dis., Tra.
NBML	Exp., Eat.
Tweet	Any plans for the weekend? How about some curry? We're opening a new curry restaurant in front of the Takaragaike baseball stadium on the 24th.

Conclusion

In this paper, we estimated the appropriate aspects of unknown tweets by introducing a feedback entropy method into a hierarchical estimation framework (HEF) for multi-label classification. Our method features two phase semi-supervised machine learning, in which many topics are extracted from a sea of tweets using an unsupervised learning model LDA. Associations among many topics and fewer aspects are built using labeled tweets. Using topics, aspects are associated with various keywords by a small set of labeled tweets. To refine the associations between topics and aspects, we evaluated the Shannon entropy of each aspect and topic. Feedback coefficients were iteratively calculated by entropy to achieve optimal associations.

To evaluate our method's effectiveness, we collected 2,390,553 tweets with the Japanese location information of "Kyoto" and prepared a small set of labeled tweets based on the classifications of three examinees. From our experimental evaluation results, our prototype system demonstrated that HEF can appropriately estimate some aspects of all the unknown tweets. Entropy feedback refines the associations between topics and aspects more than without them. We compared the results of our method and the typical methods of multi-label classification; HEF showed the highest F-measure among them. With less training data, the precision, recall, and F-measure values of the typical methods rapidly dropped; however, HEF retained its high evaluation values. Especially in F-measure, HEF usually achieved the highest score in every method. These results show that our method is effective as multi-label classification using a small labeled dataset against such short sentences as tweets. In the future, we will confirm the effectiveness of our method using other datasets, such as newspapers and blogs.

Acknowledgements

This work was supported by a Grant-in-Aid for scientific Research No.25280110.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Chang, C., and Lin, C. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):1–27.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *The Journal of Machine Learning Research* 20(3):273–297.

Domingos, P., and Pazzani, M. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *The Journal of Machine Learning Research* 29(2-3):103–130.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *The National Academy of Science* 101:5228–5235.

Hsu, C.; Chang, C.; and Lin, C. 2010. a practical guide to support vector classification. http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.

Kudo, T. 2005. Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/.

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the SIGMOD2010*, 1155–1158.

Mizunuma, Y.; Yamamoto, S.; Yamaguchi, Y.; Ikeuchi, A.; Satoh, T.; and Shimada, S. 2014. Twitter bursts: Analysis of their occurrences and classifications. In *Proceedings of the ICDS 2014*, 182–187.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press. 58.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the EMNLP 2009*, 248–256.

Riedl, M., and Biemann, C. 2012. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of the ACL 2012*, 37–42.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the WWW 2010*, 851–860.

Shannon, C. E. 1951. Prediction and entropy of printed english. *The Bell System Technical Journal* 30:50–62.

Twitter. 2014. Twitter reports fourth quarter and fiscal year 2013 results. https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321.

Wang, B.; Wang, C.; Bu, J.; Chen, C.; Zhang, W. V.; Cai, D.; and He, X. 2013. Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the WWW 2013*, 1331–1340.

Wei, Z.; Zhang, H.; Zhang, Z.; Li, W.; and Miao, D. 2011. A naive bayesian multi-label classification algorithm with application to visualize text search results. *International Journal of Advanced Intelligence* 3(2):173–188.

Yamamoto, S., and Satoh, T. 2013. Two phase extraction method for extracting real life tweets using lda. *APWeb 2013. LNCS* 7808:340–347.

Yamamoto, S., and Satoh, T. 2014. Two phase estimation method for multi-classifying real life tweets. *International Journal of Web Information Systems* 10(4):378–393.

Zhang, Y. C.; Séaghdha, D. O.; Quercia, D.; and Jambor, T. 2012. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the WSDM 2012*, 13–22.

Zhao, Z., and Mei, Q. 2013. Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the WWW 2013*, 1545–1556.

Zhao, W. X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Lim, E.-P.; and Li, X. 2011. Topical keyphrase extraction from twitter. In *Proceedings of the HLT 2011*, 379–388.