Vol.2013-MBL-68 No.1
Vol.2013-ITS-55 No.1
Vol.2013-DCC-5 No.1
2013/11/14

Activity Prediction based on both Long Term and Current Activity on Twitter

TAKUYA SHINMURA[1] YUSUKE FUKAZAWA[2]
DANDAN ZHU[3] JUN OTA[4]

***Abstract*—**we propose a method of predicting human's activity, including the location and purpose, by using Twitter posts with location information. The proposed method predicts target users' activities based on the location transition and tweet of users in the database. Concretely, we adopt both the similarity of current location and interest, and the similarity of long term interest and location to select the base user and tweet. And then, we can utilize these two baselines to predict target users' activities. We evaluate the proposed method by the following two points: one is the error range of the distance, and the other is the similarity of tweet contents. We used three months of Twitter data with location information (almost 40 mil.) as the database. The experiment results demonstrate that the prediction accuracy of the proposed method is superior to the two control groups which only consider one of the similarity of current location and interest and the similarity of long term interest and location.

Keywords*—twitter; activity prediction; data mining; social-network*

1. INTRODUCTION

1.1 Background

With the dramatically growth of diverse information, the technology that can deal with the information and process these data into tractable form is needed [1]. As an approach to these technologies, activity prediction is one of the most important research fields in computational social science. In addition, the prediction of users' destinations and activities is a crucial point in personalized recommendation and efficient navigation [2].

Generally, human's activities depend on the complexity of external environment [3]. Previous research about activity prediction covers two aspects: the destination and the purpose. For example, destination prediction by using GPS location data, purpose and destination estimation by using text mining [4], video camera [5] or behavioral pattern [6].

Due to the discreteness and diversity, the blockbuster-expanded online data need to be reorganized into optimal forms and filtered for different applications. On the other hand, the online data contain enormous amount of underlying information, and thus, they provide us excellent resources to explore valuable information [7].

1.2 Human activity

In this research, we conduct the activity prediction from two aspects: one is where people are going, which is called the destination, and the other is the purpose, that is, what people will do. We use the tweets with location information as the resource. Therefore, there are two types of data we can use: one is the text of the tweet, which is shown as "Word" in the table Ⅰ; the other is the location information shown as "Location". Besides the current information, considering the history data would contribute to the prediction accuracy, we introduce the history data about both the text and location, and the history and current are presented as "long term information" and "short term information", respectively. Therefore, the key to predict human's activity includes four factors as Table 1 shows.

Table Ⅰ: Human's activity's factors

	Word	Location
Long Term information	LW	LL
Short Term information	SW	SL

1.3 Related Works

1) Prediction by location data

The tendency to predict target user's destination [8] is to use GPS data which are continual and highly precise. By combining with Bayesian network, behavioral pattern and some other information, the accuracy of destination prediction is much improved. However, this method only considers "LL" and "SL" as shown in Table 1, but ignoring the other two factors. However, in some applications, such as recommender system, accurate prediction of target user's purpose is significant for providing appropriate services. Therefore, our method which considers both the purpose and location of users is needed.

In the research [9], it is proved that the accuracy of the destination prediction is improved when introducing the information from the friends of the target user. Therefore, we refer to this discovery and utilize the information from other users who have similar preferences to improve the prediction accuracy.

[1] Faculty of Engineering, the University of Tokyo
[2] NTT DOCOMO, INC.
[3] Faculty of Engineering, the University of Tokyo
[4] Faculty of Engineering, the University of Tokyo

2) Prediction by text data

Text data, such as blogs and posts on Social Network Service (SNS), have a large amount of information which cannot be derived from GPS data. By analyzing such kind of data, we can explore people's interests, schedules and tendencies [10]. Therefore, text data is an appropriate resource for purpose prediction.

In this research, we use "Twitter [11]" as the data resource since the data extracted from it can include both two forms, that is, the location information and the text contents. However, text data, especially the posts on SNS, include so much noise that may cause prediction accuracy reduction. So, a filter is necessary to remove meaningless words from the raw data.

Unlike the GPS data, the location information in Twitter posts is discrete, and thus, we propose a novel method to conduct destination prediction, although it will cause some loss of accuracy.

1.4 Objectives

Our final goal is to estimate target user's destination and purpose when he posts tweet with location information. The innovation is to use both the long term information and the short term information to better improve activity prediction. And the challenging points of our approach are listed as follows:

- Comprehensive utilization of the four activity factors shown in Table 1 for activity prediction;
- Exploitation of the large-scale discrete location information;
- Filter creation for removing meaningless words to improve prediction accuracy;
- Prediction of target user's destination by the discrete location data; and
- Introduction of similar user's information to improve the prediction of the target user.
-

2. METHOD

2.1 Aproach

Our research is based on the assumption that the activity of target user is similar to the past activities of similar users at the same place. The similar user means the one who has similar interests, tendencies and fields of activity.

Figure 1 is the flowchart of the proposed method. First, when target user posts his tweet with location information, we search the similar tweet posted by similar user from the past-tweet-database. And the selected tweet is named "base-tweet", and the user who posted it is called "base-user". And then, we analyze the base-tweet as well as the tweets posted by the base-user in the next hour. Finally, we work out the target user's activity and output the keywords about the purpose or destination.

We evaluated the proposed method from the purpose and destination predictions, respectively. In addition, the contributions of long term information and short term information to each prediction will be verified.

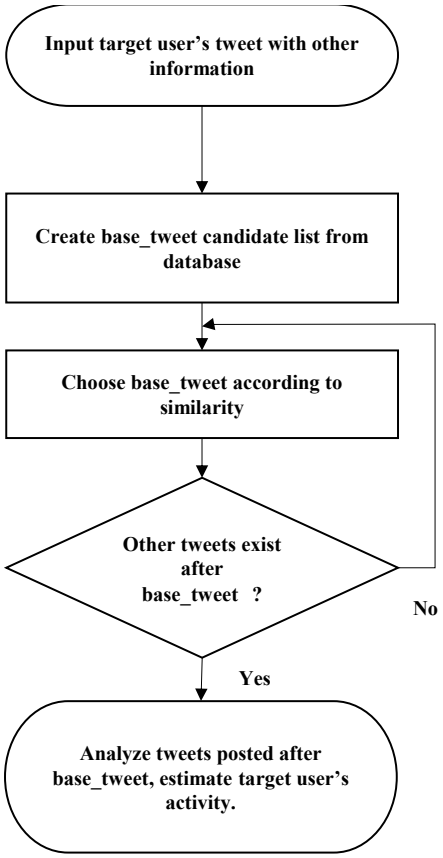


Figure 1: Flowchart of activity estimation

2.2 Database

The tweets we used for database building is described as follows:

- Tweets with location information posted between May 2011 and July 2011. The total number is about 40millions.
- Each Tweet data consists of the text of post, latitude, longitude, user ID, created time, language and other ID information for management.
- Tweets using the language of English.
- Randomly choose the target user from the database for testing.
- Remove duplicate users.
- Filter out stop words and other meaningless words.

2.3 Algorithm

We output the coordinates (latitude and longitude) of the place where the target user is going within an hour as the destination prediction. And the other output is the keywords expressing target user's purpose or destination, which is designated as the purpose prediction. For example, if a target user is hungry and posts "I am going to have a lunch", then the proposed method can work out the name of the restaurant he may go, or the name of the food he may eat within 1 hour. The reason we set the time window as one hour is that people's

schedule is planned by hour in general, and thus, when we predict the next activity of a person, one hour is appropriate.

There are two key points for the realization of the proposed method: One is the base-tweet selection, and the other is the activity prediction by analyzing the base-tweet and other tweets posted by the base-user.

3) Select base-tweet

The base-tweet is selected based on the following criterions:

- The tweet posted near the place where the target tweet was posted.
- The tweet posted by the similar user.
- The tweet which is similar to the target tweet.

In this research we designate the place where the target user posted the target tweet as "target-place".
When the target tweet is inputted, this program selects all the tweets posted within 1,000 meters of the target-place, which are defined as "candidate-tweets" and the shown as "SL" in Table 1.
And then, we create the list of similar user candidates by selecting all the users who posted these candidate-tweets in the past without duplications. Therefore, a user who has never come near the target place cannot be chosen as a similar user in this research.
Next, we need to calculate the similarity between the target user and each similar user candidate. And we defined this similarity as the "long-term-similarity".
Thereafter, we calculate the similarity between the target tweet and each candidate-tweet, which is the "current-similarity".
Finally, by combining these two similarities according to (1), we can obtain the total similarity between the target tweet and each candidate-tweet, and accordingly, the base-tweet candidate list can be created.

$$\boldsymbol{total\ similarity} = \boldsymbol{\alpha} \cdot \boldsymbol{LTS} + (\mathbf{1} - \boldsymbol{\alpha}) \cdot \boldsymbol{CS} \qquad (\mathbf{1})$$

LTS: long term similarity$(0.0 \leq LTS \leq 1.0)$ $CS: current\ similarity$ $(0.0 \leq CS \leq 1.0)$
$\alpha: the\ weight\ of\ LTS$ $(0.0 \leq \alpha \leq 1.0)$

a) Long Term Similariy

Long-term-similarity is similarity between the target user and other users who posted candidate-tweets. It considers both the past tweets and the posting place of these reference users, which can be used to derive the tweet word similarity and the location similarity, respectively. The long-term-similarity is defined by the following equation:

$$\boldsymbol{LTS} = \boldsymbol{\beta} \cdot \boldsymbol{TWS} + (\mathbf{1} - \boldsymbol{\beta}) \cdot \boldsymbol{LS} \qquad (\mathbf{2})$$

TWS: tweet word similarity$(0.0 \leq TWS \leq 1.0)$
$L: location\ similarity$ $(0.0 \leq LS \leq 1.0)$ $\beta: the\ weight\ of\ TWS$ $(0.0 \leq \beta \leq 1.0)$

Figure 2 indicates the relationship between LTS, CS, TWS and LS.

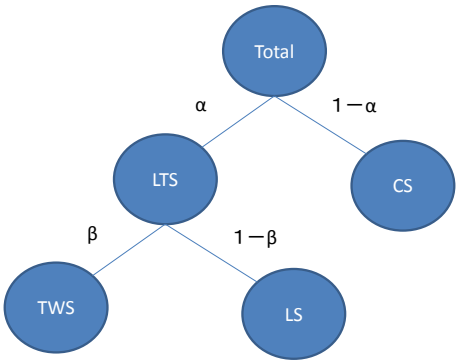


Figure 2: Relationship between parameters

A-1) TWEET WORD SIMILARITY

TWS is adopted for representing the similarity between the two tweets written by the target user and each base-user candidate. The high value of TWS indicates high similarity of the comparison users in their interests and tendencies. Each user is expressed as an N-dimensional status vector in the calculation of TWS whose vector components are the text words.
In this process, we extract the words used in all the tweets posted by the target user and the base-user candidates without duplications. First, we create the Hash Map vector form for each user, whose components are the extracted words. The size of this vector depends on the number of the candidate users' tweets. For example, when we choose a target tweet posted in New York, there are approximately 60,000 users around the posting place of the target tweet, making the 60, 000-dimension vector.
And then, we count the appearance of each word obtained by the extraction, and use these values to assign the status vector which represents the tweet contents of each user.
Finally, the TWS between the target user and each base-user candidate is defined by using the created status vector and the cosine method. This method is often used in recommendation collaborative filtering [12], since it can provide reliable similarity between users.

$$\mathbf{TWS} = \cos(\vec{\boldsymbol{A}}, \vec{\boldsymbol{B}}) = \frac{\vec{\boldsymbol{A}} \cdot \vec{\boldsymbol{B}}}{|\vec{\boldsymbol{A}}| \cdot |\vec{\boldsymbol{B}}|} \qquad (\mathbf{3})$$

A: target user's word vector
B: a candidate user's word vector

A-2) LOCATION SIMILARITY

LS describes the similarity of the activity locations, and a higher LS value indicates that the target user and the candidate user

share more in where they live and where they usually go. Each user is represented as 26-dimensional status vector to calculate LS. And each component of the vector is a given-location ID number which indexes a rectangular area around the posting place of the target tweet. Figure 3 gives an example to show the detailed indexing, in which target tweet is posted in No.13 area. From the fig. 3, we can see that the map zooms and centers on the location of the target-place, and selects the 25×25 kilometers square region as the study subject. Then, the selected square region is divided into 25 square blocks, each of which is 5 kilometers on a side, plus an extra block for other places beyond the chosen region. And the target-place is located in the central block which is marked as No.13.

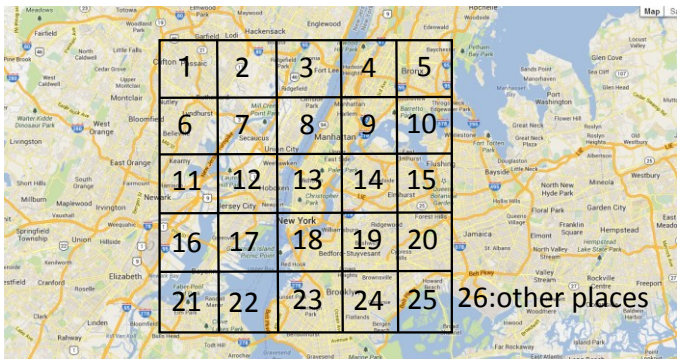


Figure 3: Map separation image

After that, we can use this built map to analyze the location information of each candidate user and assign the integer values to the LS vector. For a given candidate user, we need to check her/his tweets posted in all the 26 block areas, count the appearances of the tweets posed in each block area, and the obtained counts are the assigned values for the LS vector's according components. For example, a user posted all her/his 10 tweets within 2.25 kilometer distance from the target-place, and then, in her/his LS vector, the value of the No.13 area component is 10, while the others are 0.
After enabling the LS vectors, we can calculate the LS between the target user and other users by the following cosine method:

$$\mathbf{LS} = \cos(\vec{C}, \vec{D}) = \frac{\vec{C} \cdot \vec{D}}{|\vec{C}| \cdot |\vec{D}|} \qquad (4)$$

C: target user's location vector
D: each candidate user's location vector

b) Current Similarity

Current Similarity indicates the similarity of the contents between the target-tweet and each candidate-tweet. Therefore, calculation method is similar to the TWS. Each tweet is represented as an N-dimensional status vector whose components are the words used in the tweet, and the according assignment values are the occurrences of these words.

The formula of the CS between the target-tweet and each candidate-tweet is given as follows:

$$\mathbf{CS} = \cos(\vec{E}, \vec{F}) = \frac{\vec{E} \cdot \vec{F}}{|E| \cdot |\vec{F}|} \qquad (5)$$

E: target-tweet word vector
F: each candidate-tweet's word vector

By using the LTS and the CS, we can calculate the total similarity between the target-tweet and each candidate-tweet. And then, we sort the base-tweet candidate list by the total similarity in descending order. Therefore, we choose the top-listed tweet as the base-tweet. However, it can happen that there is no followed tweet after the base-tweet posted by the base-user. In this case, we choose the following top-listed tweet as the base-tweet.

4) Activity Prediction of the target user

After the identification of the base-tweet, we can conduct the activity prediction by analyzing the tweets posted by the base-user in the next hour of the base-tweet posting.

a) Destination Estimation

We use the location information of the base-user's next-hour tweets to estimate the destination of the target-user. By extract the latitude and longitude data from these tweets, we can figure out the average latitude and longitude. A simple method is that we just use these average values for the answer to the destination estimation. As we mentioned before, we have limited the estimation in one-hour time window, and it is enough for estimation since a lot of users post only one tweet within an hour. And even if people do post more than one tweets during that time, it is reasonable to believe that all these tweets are posted around the one-hour-destination unless the user is in the course of long-traveling.
Therefore, we output the average latitude and longitude as the destination estimation of the target user.

b) Purpose Estimation

The text part of the next-hour tweets, which is written by the users, is utilized for purpose estimation. To find the key words of users' purpose, we introduce the technology of Term Frequency Inverse Document Frequency [11] which is represented as "tf-idf" in (6). It can quantitatively evaluate the contribution of each word in a document. The value of tf-idf is higher when the word occurs more frequently in the document but less used in other documents. We can calculate the tf-idf of each word by the following equation:

$$\begin{aligned} \mathbf{tf\ df} &= \mathbf{tf} \cdot \mathbf{idf} \\ \mathbf{tf_{i,j}} &= \frac{n_{i,j}}{\sum_k n_{k,j}} \\ \mathbf{idf_i} &= \log \frac{D}{d_i} \end{aligned} \qquad (6)$$

$n_{i,j}$: *the number of word* "i" *in document* "j"
D : the number of total documents
d_i : *the number of documents which contain word* "*i*"

According to the tf-idf values of words, we choose the highest one as the key word to describe the purpose of the target user.

3. EVALUATIONS

To verify the performance of the proposed method in activity prediction, we explore the contributions of the related influences to the prediction accuracy. There are four factors needed for activity prediction: LTS, CS, TWS and LS. And according the definition formulas (1) and (2), each of them has a contribution weight, that is, α , 1- α , β and 1- β , respectively. Because the weight of LTS is actually the complementation of the CS weight, and so do the TWS weight to the LS weight. Therefore, we only need to deal with the two parameters, that is, the weights of LTS and TWS, and treat them as variables, respectively, to figure out how these two parameters affect the prediction performance.

3.1 Details of implenmentation

The quantitative evaluation is based on the following key points:

- The evaluation of the activity prediction is promoted in two ways: destination prediction and purpose prediction.
- Few tweets around the target area may cause null value for purpose prediction, and in that case, the purpose output is excluded.
- The tweets posted by bot user may cause the predicted destination ridiculous far from the posting location, which is impossible for people to reach within one hour, and thus, such results need to be removed.
- Varying the LTS and TWS, respectively, by using 0.1 as the sample step, we conducted experiments 121 times for each target-tweet in total.
- The target tweets for testing are randomly selected, and the number of these tweets is about 1000.

3.2 Evaluation on Destination Prediction

We introduce the error range to evaluate the performance of the destination prediction. The error range is defined as the distance between the predicted destination and the golden answer. Therefore, the lower this value is, the better the prediction accuracy is. As for the creation of the golden answer to the destination prediction, we select the next-hour tweets posted by the target user, extract the location information from them, and use the average coordinate data as the golden answer.

3.3 Evaluation on Purpose Prediction

We use the similarity of the next-hour tweet contents between the target user and the base user to measure the purpose prediction accuracy. The time references are the target tweet posting time and the base tweet posting time, respectively. And the calculation method is similar to the TWS and the CS. The higher value indicates higher prediction accuracy.

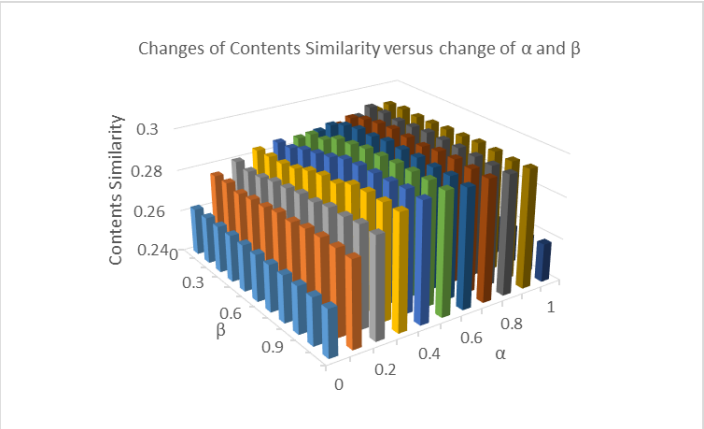
It is notable that the output keyword has been removed from each tweet in the process of the evaluation.

4. RESULTS

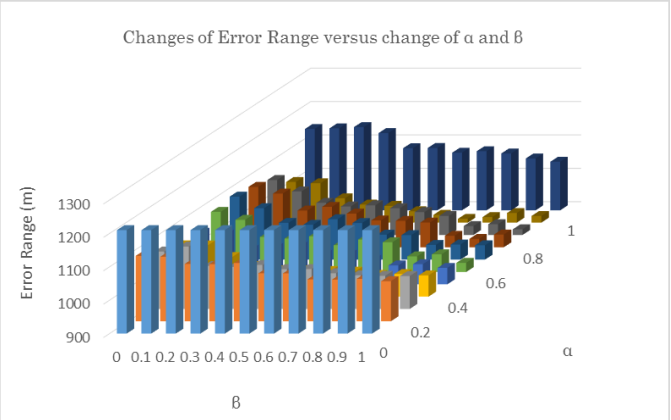
The experimental results are shown in fig.4. From the results, we can see that the optimal prediction accuracy of the purpose is obtained when the weight of LTS α is 0.7, and the weight of TWS β is 0.8, which reaches 29.84%. And for the destination prediction, when α is 0.9 and β is 0.7, the error range attains the minimum, that is, 911.40 meters.

The results demonstrate that the LTS and the TWS have far greater impacts on the prediction accuracy than the CS and the LS. This discovery means that long term information plays a more important role than short term information in the activity prediction, and people's future activities are more relevant to their interests and tendencies than the locations.

And taking both of the destination and purpose prediction into account, we can tell that the best prediction performance comes when the value for α ranges between 0.7 to 0.9 and the value for β ranges between 0.7 to 0.8.



(a) Relation between the Contents Similarity and (α , β)



(b) Relation between the Error Range and (α , β)

Figure 4. Evaluation Results

In addition, in Fig. 5, we present the optimal parameters obtained by evaluations on the purpose and destination prediction, respectively.

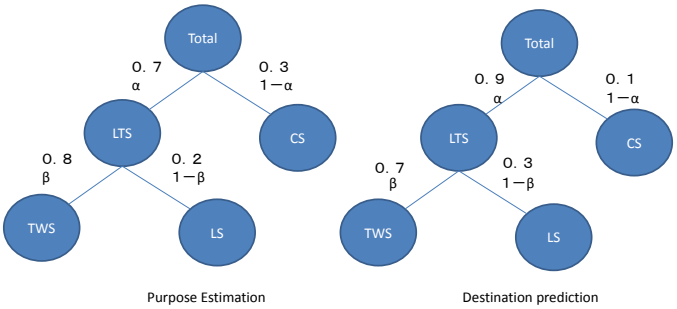


Figure 5: Optimal parameters

5. DISCUSSION

This result proved that Long Term Similarity greatly affects the accuracy of human's activity prediction. As an effective factor of Long Term Similarity, the Tweet Word Similarity is more important than Location Similarity. However, we can tell that even for the destination prediction, the contribution of the Location Similarity is weak. It may because the proposed method cannot completely filter the bot tweets, causing the noisy for the destination prediction.

It is important to note that our method is not suitable for all-round activity prediction. So, we need to classify human's activities and improve our method to fit to the all-round activity prediction [13].

6. CONCLUSION

In this research, we proposed a novel method to predict users' activity, discussed the contributing factor in the prediction accuracy, and worked out the optimal parameters for the accuracy. This result shows that long term information is more important than short term information when we predict human's activity and their posted contents effect the long term factor more than the activity area. However, when we predict human's activity, we cannot ignore other factors in Table 1. It proved that human's activity depends much more on his long term intention than his intention at that time.

In future, we are planning to create an android application based on the proposed method, and conduct real-life tests for comprehensively evaluate it. So, before the application, we should refine our method to improve the prediction accuracy, and do large-scale experiments to verify the feasibility. And we hope that this technology will bring benefits to various kinds of services, such as the collaboration of e-commerce web sites, recommendations and social network services.

References

[1] Ministry of Economy, Trade and Industry, Information Grand Voyage Project, available: http://www.meti.go.jp , Accessed 2013 July 13.

[2] Yusuke Fukazawa, Jun Ota, Automatic task-based profile representation for content-based recommendation, KES Journal 16(4), pp247-260, 2012.

[3] Herbert A.Simon, The Science of The Artificial, Personal Media, 1999.

[4] Asuka Sumida, Gen Hattori and Tomohiro Ono, Trial manufacture of system that can estimate trouble attributed to individual activity by using Twitter,The Association for Natural Language Processing, pp456-459, 2011.

[5] M.S.Ryoo, Human Activity Prediction:Early Recognition of Ongoing Activities from Streaming Videos,ICCV,pp286-299, 2011.

[6] Fumitaka Nakahara and Takahiro Murakami, A Destination Prediction Method Based on Behavioral Pattern Analysis of Nonperiodic Position Logs, ICMU, pp 32-39, 2012.

[7] Thorsten Joachims and Filip Radlinski, Search Engines that Learn from Implicit Feedback, IEEE Computer Society, pp34-40,2007

[8] Naoharu Yamada, Yoshinori Isoda, Masateru Minami and Hiroyuki Morikawa, Incremental Route Refinement for GPS-enabled Cellular Phones, ICMU, pp87-93, 2010.

[9] Manlio De Domenico, Antonio Lima and Mirco Musolesi, Interdependence and Predictability of Human Mobility and Social Interactions, eprint arXiv:1210.2376, 2012.

[10] Dandan Zhu, Yusuke Fukazawa, Eleftherios Karapetsas, Jun Ota, Intuitive Topic Discovery by Incorporating Word-Pair's Connection Into LDA, Web Intelligence 2012, pp303-310,2012.

[11] Twitter, available: https://twitter.com/, Accessed 2013 July 13

[12] Greg Linden, Brent Smith and Jeremy York, Amazon.com Recommendations Item-to-Item Collaborative Filtering, IEEE internet computing, pp76-80, 2003.

[13] Salton G, Buckley C, "Term-weighting approaches in automatic text retrieval", Information Processing and Management 24 (5), pp513-523, 1988.

[14] Statistics Japan, Survey of life of the people, available: http://www.stat.go.jp/, Accessed 2013 July 13.