Recommending Fresh URLs Using Twitter Lists

Yuto Yamaguchi, Toshiyuki Amagasa and **Hiroyuki Kitagawa**
University of Tsukuba, Japan
yuto_ymgc@kde.cs.tsukuba.ac.jp, {amagasa, kitagawa}@cs.tsukuba.ac.jp

Abstract

Recommender systems for social media have attracted considerable attentions due to its inherent features, such as a huge amount of information, social networks, and real-time features. In microblogs, which have been recognized as one of the most popular social media, most of URLs posted by users are considered to be fresh (i.e., shortly after creation). Hence, it is important to recommend URLs in microblogs for appropriate users because users become able to obtain such fresh URLs immediately. In this paper, we propose a URL recommender system using Twitter user lists. Twitter user list is the official functionality to group users into a list along with the name of it. Since it is expected that the members of a list (i.e., users included in the list) have similar characteristics, we utilize this feature to capture the user interests. Experimental results show that our proposed method achieves higher effectiveness than other methods based on the follow-followed network which does not offer user interests explicitly.

Introduction

This paper focuses on recommending URLs posted as parts of tweets for Twitter users. Twitter is one of the most notable social media providing us with a vast amount of real time information. URLs in Twitter are valuable because of their interesting features. First, Twitter users often include URLs in their *tweets*. By including URLs in their tweets, users can share resources in the web such as breaking news stories with their *followers*. According to our preliminary experiment, it is revealed that about 14% of tweets contain at least one URL, resulting in that a considerable number of URLs are posted and shared in Twitter. Second, these URLs are checked by users before included in tweets whether they are worth to share or not. This results in that shared URLs are of high quality (i.e., useful, interesting, or informative). Third, most of URLs posted in Twitter are *fresh* (Dong et al. 2010). In other words, these URLs are posted within short time periods after the respective web pages are created. Therefore, we can obtain fresh information from Twitter.

Most work to address the problem of URL recommendation in Twitter are based on contents of web pages and tweets, or topology of social networks (Kywe, Lim, and Zhu 2012). In the former case, recommender systems analyze the contents of web pages and tweets to capture the user interests and the topic of URLs. However, since web pages are of various types (e.g., texts, videos, and photos), it is quite difficult for contents based recommender systems to analyze such various types of contents. To deal with different types of web pages, we would rather choose collaborative filtering approaches, which corresponds to the latter case. In Twitter, users can *follow* other users to receive their tweets, which forms the *follow-followed network*. Basically, users follow other users based on their interests. However, the follow-followed network does not represent user interests explicitly. Thus, if we want to capture interests of users, we may need aother source of information.

In this work, Twitter users are potentially classified into three types, namely, information sources, information seekers, and communication users. Information sources would rather transmit information than collect information. These users mainly post tweets about topics they are interested in. We call this topic *user expertise topic*. If a URL is posted by many users whose expertise topic is t, we can infer that the URL is likely useful in topic t. Information seekers, the second type of users, would like to collect information about the topic they are interested in. We call this topic *user interest topic*. This type of users may potentially have information needs about their interest topics, which motivates us to develop the recommender system targetting this type of users in Twitter. Communication users, the third type, do not have clear expertise or interest topics. These users use Twitter just as a communication tool. So we do not focus on these users. If we can infer user expertise topics and user interest topics, it becomes possible to capture the topics of URLs, thereby enabling recommendation for appropriate users who are interested in respective topics.

To capture user expertise topics and user interest topics, we utilize *Twitter user lists*. Twitter user list is the official functionality to group users into a list and specify the name of it. All users are allowed to make and share lists with other users. In our previous work (Yamaguchi, Amagasa, and Kitagawa 2012), we revealed that users who often transmit information about a topic tend to be included in many lists about that topic by other users. For example, a user who often transmits information about tennis is likely to be included into many lists named *tennis*. In this paper, according to this observation, we set three assumptions; 1) users

with expertise topic t tend to be included in lists about topic t, 2) users with interest topic t tend to follow users whose expertise topics cover t, and 3) URLs posted by many users whose expertise topic is t are useful on topic t.

Contributions in this paper are as follows.

- *Proposing a URL recommender system utilizing Twitter user lists:* The proposed method has three components, capturing user expertise topics, user interest topics, and topic-wise usefulness of URLs.
- *Demonstrating our method works well in the case recommending URLs for information seekers:* Experimental results show that our method works well in the context of recommending URLs for information seekers, because their clear interests can be inferred by using Twitter lists.

Related Work

There are several works proposing URL recommender systems utilizing Twitter. (Abel et al. 2011) and (Phelan, McCarthy, and Smyth 2009) both proposed a method to recommend news stories for Twitter users. These methods analyze the contents of news articles or the contents of tweets to capture the users' interests. (Chen et al. 2010) proposed several kind of URL recommending methods in Twitter and compared them experimentally. (Hannon, Bennett, and Smyth 2010) proposed a method to recommend users who often transmit information about the topic the target user is interested in.

There are two main differences between these works and ours. First, our method incorporates the user expertise topics. This enables our method to measure topic-wise usefulness of URLs by aggregating the expertise topics of users who posted the URL. Second, our method does not use any contents of tweets and web pages. Hence our method is based only on the graph structure.

In addition to these researches, (Dong et al. 2010) proposed a method to rank URLs posted in Twitter to obtain useful and fresh URLs. They reported that fresh URLs can be more effectively collected from Twitter than crawling the entire web. (Kywe, Lim, and Zhu 2012) surveyed wide range of recommender systems in Twitter. (García-Silva et al. 2012) and (Kim et al. 2010) investigated Twitter lists for user characterization. Both the researches reported that the terms in list names describe user interests or characteristics appropriately.

Proposed System

The procedure of our proposed system is as follows.

1. *Estimating user expertise topics and user interest topics:* The topics are estimated by membership relations between lists and users, and follow relations between users.
2. *Constructing candidate user sets:* Candidate user sets consist of users who have high probability to post the URL the target user may be interested in. Only URLs posted by users in the candidate user set are candidates for recommendation.
3. *Calculating recommending scores:* Recommending scores, defined between the target user and a URL, indicate whether the URL should be recommended for the target user or not.

In the following subsections, we detail these steps.

Estimating User Expertise Topics

Firstly, we estimate user expertise topics using lists. The topic distribution of lists in which a user is included represents the distribution of expertise topics of the user. For example, consider a user who is included in 100 lists, 70 of them are about politics, 20 of them are about sports, and 10 of them are about music. The topic distribution of these lists tells us that this user mainly posts tweets about politics but sometimes posts tweets about sports or music. According to our previous work (Yamaguchi, Amagasa, and Kitagawa 2012), this distribution can be captured by the term distribution of list names. In case of the above example, the names of 70 lists about politics probably contain terms related to politics. Hence we can estimate user expertise topics by the term distribution of list names.

Let u be each user and t be each term in list names. User expertise topic e_{ut}, which denotes the term distribution of u, is calculated by Equation 1.

$$e_{ut} = \frac{\sum_{l \in lists(u)} \delta_{lt}}{\sum_u \sum_{l \in list(u)} \delta_{lt}} \quad (1)$$

where $lists(u)$ denotes the set of lists u is included in, δ_{lt} takes 1 if the name of list l contains t; otherwise it takes 0. When e_{ut} is large, it can be considered that one of u's expertise topics is t, and u tends to post tweets about t.

Estimating User Interest Topics

Next, we estimate user interest topics using calculated expertise topics and follow relationships between users. We aggregate the expertise topics of users followed by the target user. Hence, user interest topics are also denoted as the term distribution. User interest topic p_{ut} is calculated by Equation 2.

$$p_{ut} = \frac{idf(t) \sum_{v \in follow(u)} e_{ut}}{\sum_t idf(t) \sum_{v \in follow(u)} e_{lt}},\ idf(t) = \log \frac{|L|}{N_t} \quad (2)$$

$follow(u)$ denotes the set of users followed by u, and $idf(t)$ denotes the value of inverse document frequency, where L denotes the set of all lists and N_t denotes the number of lists contain t in their names. The larger p_{ut} is, the more strongly u is interested in t.

Constructing Candidate User Sets

For each target user, we construct a candidate user set. Candidate user sets consist of users who are the most likely to post URLs the target user is interested in. In our proposed system, only URLs posted by users in candidate user sets have chances to be recommended for the target user. Constructing candidate user sets reduces computational costs and noises, because we need not deal with all URLs.

Candidate user sets are constructed as follows. First, for each target user, we extract top k_T terms whose value of p_{ut} is the largest, dentoted as C_u^T. These terms belonging to C_u^T

Table 1: Details of the dataset.

	size		size
$\|U\|$	501,777	$\|W\|$	8,778,855
$\|L\|$	1,179,129	$\|T_w\|$	12,766,847
$\|U_M\|$	3,873,979	$\|T_e\|$	56,819

show the interest topics of u appropriately. Second, for each term belonging to C_u^T, we extract top k_U users whose value of e_{ut} is the largest, denoted as C_u^U. These users are the most likely to post URLs that u is interested in. k_T and k_U, which are parameters, are determined experimentally.

Calculating Recommending Scores

For each target user u, the recommending score s_{uw}, which determines whether URL w is to be recommended for u or not, is calculated by Equation 3.

$$s_{uw} = \sum_{v \in C_u^U} \sum_{t \in C_u^T} p_{ut} \cdot e_{vt} \cdot \theta_{vw} \quad (3)$$

where θ_{vw} denotes whether user v posts w or not, which takes 1 if v posts w, otherwise takes 0. Recommending scores are high if the corresponding URL is posted by many users whose expertise topics are similar to u's interest topic. Top N URLs with the highest recommending scores are recommended for target users.

Experiments

In this section, we discuss our experiment which is conducted to verify whether our method works well or not. In summary, our method could infer user interest topics of information seekers and appropriately recommend URLs for them. While for communication users, a method based purely on the follow-followed network works better, which is probably because of the well-known *homophily effect*.

Experimental Data

We collected the data of users U, lists L, follow relationships between users, and membership relationships between lists and users U_M using Twitter API. Terms contained in the name of lists in L are extracted, denoted by T_e. Tweets containing at least one URL are collected from Nov. 15th to 18th in 2011 using Twitter Streaming API. The set of collected tweets is denoted by T_w, and the set of URLs contained in that tweets is denoted by W. Details of the constructed dataset are shown in Table 1.

Compared Methods

In this experiment, three recommending methods are compared in terms of the recommendation accuracy. *Proposed* is our proposed method with parameters $k_T = k_U = 30$, which are defined by preliminary experimental results. Preliminary results indicated that even if these parameters are set to larger than 30, recommendation results are not improved but the compotational cost becomes expensive.

Follow is the method purely based on the follow-followed network. This method is based on the observation that user u and user $v \in follow(u)$ have similar interests (Chen et al. 2010). Hence, u is also likely to be interested in information transmitted by user $z \in follow(v)$ who is within two hops from u in the follow-followed network. The set of users within two hops from u are denoted as $FF(u)$. If user $z \in FF(u)$ is followed by a lot of users in $FF(u)$, z is more likely to post a URL u is interested in, because a lot of users who have similar interests as u follow z. Based on this idea, recommending scores of *Follow* are calculated by $s_{uw} = \sum_{z \in FF} trust(u, z) \cdot \theta_{zw}$, where $trust(u, z)$ denotes the number of z's followers within $FF(u)$.

Popular recommends the most popular URLs for all target users uniformly. Recommending scores of this method are simply calculated by $s_{uw} = \sum_{v \in U} \theta_{vw}$.

Details of Examinees

In this experiment, using above three compared methods, we recommend URLs to 12 examinees, who are graduate students majoring computer science. We can consider that our method works well for information seekers, because unlike other methods, our method can explicitly infer user interest topics. To verify this, we divided 12 examinees into two groups; the first group includes information seekers, and the second group includes others. Information seekers tend to follow famous and useful users who often transmit useful information. These users are likely to have a lot of followers. Therefore, we can determine examinees who follow a lot of these users as information seekers.

We simply divided examinees into two groups as follows. First, for each examinee, we calculate the number of followers of users who are followed by the examinee. Then, top 6 examinees with the large calculated number are assigned to the first group, and the other users are assigned to the second group. Consequently, examinees in the first group follow a lot of famous and useful users, that is, these examinees can more or less be recognized as information seekers. Note that although there are several kind of methods to extract useful users (Weng et al. 2010) (Leavitt et al. 2009), we adopt the most simple metric.

Evaluation Process

Evaluation process of this experiment is as follows. First, for each examinee, recommending scores of all candidate URLs are calculated by three compared methods. Second, for each examinee, and for each compared method, top 20 URLs that have the highest recommending scores are recommended. Third, examinees evaluate each URLs which are recommended to themselves whether they are interested in these URLs or not. Each URL is evaluated by examinees in levels from 1 to 5, where 1 means *completely not interesting*, and 5 means *very interesting*. If the examinee assigns a URL 3 or higher score, we judge the URL to be relevant to the examinee. The result of this evaluation is calculated using Precision@k. Precision@k is the average of the precisions of top k URLs recommended for the corresponding examinee. Figure 1 and Figure 2 show the result of the examinee in the first and the second group, respectively.

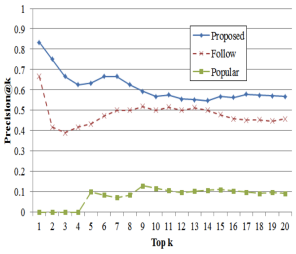

Figure 1: Precision@k of the first group.

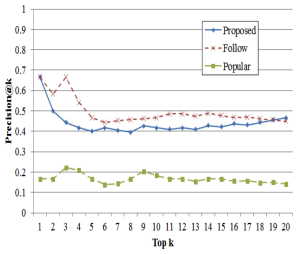

Figure 2: Precision@k of the second group.

Results

In Figures 1 and 2, both *Proposed* and *Follow* achieve much higher precisions than that of *Popular*. This indicates that recommending URLs not considering user interests does not make sense even if popular URLs are recommended.

In Figure 1, for examinees in the first group, *Proposed* outperforms *Follow*, indicating that *Proposed* works well for information seekers. While in Figure 2, for examinees in the second group, *Follow* shows higher precision than that of *Proposed*, indicating that *Proposed* fails to recommend appropriate URLs for users who do not have purpose to collect information from Twitter. In the rest of this section, we discuss the results of *Proposed* and *Follow*.

It is considered that *Follow* works well, naturally, when examinees have similar interest topics with their friends. Most of examinees in this experiment are, say, ordinary users who often communicate with their friends using Twitter (i.e., communication users). This results in that *Follow* achieves higher precision than our proposed method in the case for examinees in the second group, because friends tend to have similar interests. This is well known as *homophilly effect* (Bisgin, Agarwal, and Xu 2012) (Kwak et al. 2010).

While in the case of the first group, in Figure 1, *Proposed* works well for information seekers. The reason why this result comes out can be interpreted two ways. First, the same as our hypothesis, most of examinees in the first group are information seekers who have clear interests to collect information. Indeed, these examinees follow users whose major expertise topics are similar and consistent. Second, useful users, followed by a lot of users, are also included in a lot of lists. This means that useful users have much information to estimate their expertise topics. Consequently, useful users' expertise topics are accurately estimated, and then, examinees' interest topics are also accurately estimated because interest topics are calculated aggregating expertise topics.

From these results and discussions, it is shown that *Follow* that utilizes homophily effect works well for communication users, and *Proposed* that utilizes Twitter user lists and deals with user expertise topics and interest topics works well for information seekers.

Conclusion

In this paper, we proposed a URL recommender system for Twitter users utilizing Twitter user lists. URLs posted in Twitter are mostly fresh. Hence, our method can recommend these fresh URLs. Proposed method estimates user expertise topics and user interest topics based on the follow relationships between users and the membership relationships between lists and users, and then measures topic-wise usefulness of URLs. Experimental results show that our proposed method achieves higher precision in the context of recommending URLs for information seekers than that of other recommending methods including a method based purely on the follow-followed network.

Acknowledgement

This work has been supported in part by JSPS KAKENHI, Grant-in-Aid for JSPS Fellows #242322.

References

Abel, F.; Gao, Q.; Houben, G.-J.; and Tao, K. 2011. Analyzing user modeling on twitter for personalized news recommendations. In *UMAP*, 1–12.

Bisgin, H.; Agarwal, N.; and Xu, X. 2012. A study of homophily on social media. *World Wide Web* 15(2):213–232.

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M.; and Chi, E. 2010. Short and tweet: experiments on recommending content from information streams. In *CHI*, 1185–1194.

Dong, A.; Zhang, R.; Kolari, P.; Bai, J.; Diaz, F.; Chang, Y.; Zheng, Z.; and Zha, H. 2010. Time is of the essence: improving recency ranking using twitter data. In *WWW*, 331–340.

García-Silva, A.; Kang, J.-H.; Lerman, K.; and Corcho, Ó. 2012. Characterising emergent semantics in twitter lists. In *ESWC*, 530–544.

Hannon, J.; Bennett, M.; and Smyth, B. 2010. Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys*, 199–206.

Kim, D.; Jo, Y.; Moon, I.-C.; and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. *In Workshop on Microblogging at the CHI 2010*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. B. 2010. What is twitter, a social network or a news media? In *WWW*, 591–600.

Kywe, S. M.; Lim, E.-P.; and Zhu, F. 2012. A survey of recommender systems in twitter. In *SocInfo*, 420–433.

Leavitt, A.; Burchard, E.; Fisher, D.; and Gilbert, S. 2009. The influentials: New approaches for analyzing influence on twitter. *a publication of the Web Ecology project*.

Phelan, O.; McCarthy, K.; and Smyth, B. 2009. Using twitter to recommend real-time topical news. In *RecSys*, 385–388.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 261–270.

Yamaguchi, Y.; Amagasa, T.; and Kitagawa, H. 2012. Tagging users based on twitter lists. *Int. J. Web Eng. Technol.* 7(3):273–298.