Characterizing User Interest Using Heterogeneous Media

Jonghyun Han
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
jhan@gist.ac.kr

Hyunju Lee
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
hyunjulee@gist.ac.kr

ABSTRACT

It is often hard to accurately estimate interests of social media users because their messages do not have additional information, such as a category. In this paper, we propose an approach that estimates user interest from social media to provide personalized services. Our approach employs heterogeneous media to map social media onto categories. To describe the categories, we propose a hybrid method that integrates a topic model with TF-ICF for extracting both explicitly presented and implicitly latent features. Our evaluation result shows that it gives the highest performance, compared to other approaches. Thus, we expect that the proposed approach is helpful in advancing personalization of social media services.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*User-centered design*

Keywords

User interest; Social media; Topic modeling

1. INTRODUCTION

With the increasing popularity of social media services, such as Twitter and Facebook, users of these services obtain information from their friends' brief text messages, images, and videos. They follow their acquaintances [2] or other users having similar interests with them. Users usually try to find friends who post information relevant to their interests. From the perspective of social media services, in order to provide them with user-preferred information, it becomes important to understand their interests [3]. Thus, estimating interests of users is necessary for recommending user-preferred content.

In this paper, we propose an approach that estimates interests of social media users to provide user-preferred content. We express user interest in terms of probabilities that a user has interest in categories. Since there is no information related to categories in social media, we employ traditional news media containing news articles classified into categories. Because these media are heterogeneous, it is necessary to take features of news media and social media into consideration. We extract features of the categories of news media by a hybrid method that exploits both explicitly presented term vectors and implicitly latent topic distributions. The method integrates user interests estimated by a TF-ICF approach for explicit features and a topic modeling approach for latent features. Then, we find a user's interesting categories through comparing features of the news categories and features of personal social media. We verified our proposed approach by measuring a similarity between estimated user interest by our approach and manually labeled user interest by annotators. According to our evaluation result, it is able to estimate user interest that is similar to the labeled interest from social media.

2. ESTIMATION OF USER INTEREST

In this paper, we propose a method to estimate user interest from personal social media. We express user interest in terms of categories' probabilities, which show how much a user has interest in those categories. User interest is defined as $[c_1, c_2, \cdots, c_n]^T$, where c_i represents the probability of the i-th category and n is the number of categories. We take advantage of news media that contains news articles classified into categories, because there is no information related to categories in social media. We gather news articles from a news portal site that classifies the articles into the categories. News articles posted each day in each category are gathered and regarded as a document. We extract features of the categories from the news articles using two methods; One is based on a bag-of-words approach and the other is based on a topic modeling approach.

We first extract explicitly presented features of categories. We define and use inverse category frequency (ICF) as a measure of whether a term is common or rare across all categories. Because some important terms, which frequently appear in all documents belonging to a specific category but rarely appear in other categories' documents, have low IDF values, we use ICF instead of IDF. After accumulating term counts of documents in each category and removing rarely appeared terms, ICF of a term is computed by dividing the total number of categories by the number of categories containing the term, and taking the logarithm of that quotient. Finally, we compute the product of term frequency and ICF (TF-ICF) for the significance of a term. The value of a term's TF-ICF represents features of categories.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577346.

Table 1: The most significant terms, extracted by TF-ICF and topic modeling, to describe categories

Category	TF-ICF based Approach	Topic Model-based Approach	
	Significant Terms	Topic	Significant Terms
politics	assemblyman, Geun-hye Park, president	#14	candidate, assemblyman, election, president
		#23	North Korea, national security, launch
economy	deal, stock, price, quarter, bank, increase	#07	finance, loan, bank, CEO, insurance, card
society	prosecution, police, investigation, doctor	#24	union, strike, worker, labor, wage, income
life	brand, skin, color, health, exhibition	#13	brand, skin, color, style, coffee, product
world	Obama, demonstration, Israel, president	#04	Obama, president, premier, Israel, Iran, Egypt
IT/science	game, smartphone, Samsung, release	#15	query, Apple, mobile, Samsung, Google
		#05	security, solution, data, system, mobile, Apple
entertainment	actor, drama, member, appearance	#17	movie, art work, show, audience, writer, actor
sports	league, hit, run, football, team, homerun	#22	hit, homerun, Shin-soo Choo, pitcher, hitter
		#08	football, league, coach, player, second half

To extract implicitly latent features of categories, we also propose a method that estimates user interest using a topic modeling approach. We use LDA [1] as a topic model to exploit latent topics of categories. Some terms, frequently presented in news articles – e.g., 'reporter' or 'announcement' – make unimportant topics and inaccurate the order of terms' significances in a topic. To avoid these problems, we generate documents for a topic modeling through using TF-ICF instead of a bag-of-words model. We perform LDA to get topic distributions of the generated documents. Then, the topic distributions of documents, which belong to a same category, are integrated into topic distributions of the category. The topic distributions of a category c, θ_c, represent how much the category is related to topics.

We take advantage of social media text messages posted by a user to estimate his/her interest. Because social media does not contain category information, we compare the messages with the extracted features of categories obtained by the two proposed methods. First, we measure similarity between TF-ICF values of a user's messages and TF-ICF values of each category. The similarity represents a probability that a user has interest in each category. Second, we get a user's topic distribution θ_u from the messages using a topic model trained by news articles. We measure similarity between θ_u and topic distributions θ_c of each category. It represents user interest estimated by topic modeling. Finally, we integrate user interests estimated by TF-ICF and topic modeling for considering both explicitly presented features and implicitly latent topics.

In order to verify the accuracy of interest estimated by our approach, we compared the estimated interest with interest manually labeled by two recruited human annotators. In this evaluation, we used 59 news categories, 4,764K news articles, and 22,723 messages posted by 28 Twitter users. Table 1 shows significant terms, which describe the categories, extracted by our approach. Figure 1 shows the result of our evaluation including comparisons with other approaches, such as bag-of-words, TF-ICF, and LDA. As shown in Figure 1, the proposed approach gave the highest performance, compared to the other approaches. The TF-ICF based approach performed better than the topic modeling. When we used our hybrid approach (empirically derived weight: 26% topic model, 74% TF-ICF), we estimated much more accurate interests than the other approaches. According to our evaluation result, we can find out that our approach is able to estimate user interest using personal social media.

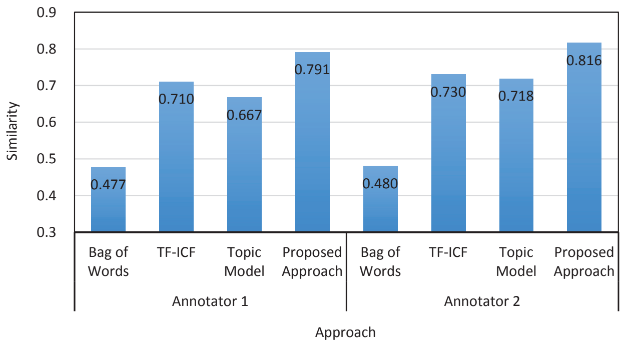

Figure 1: The evaluation result showing the similarity between estimated interest by each approach and labeled interest by two annotators.

3. DISCUSSION

In this paper, we proposed an approach that estimates user interest from social media. We employed traditional news media and utilized their category information. The approach extracts explicitly or implicitly presented features of categories through the TF-ICF method and the topic modeling method. It measures similarities between a user's social media messages and news categories using the extracted features. According to our evaluation result, estimated interest by our approach is more similar with labeled interest by annotators than those of other approaches. Future works include developing methods that enhance estimation of user interest and consider other factors of social media, such as friendship. Also, it is necessary to develop a method, which avoids problems caused by heterogeneous data sources, such as abbreviations or nicknames.

4. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] J. Han and H. Lee. Analyzing social media friendship for personalization. In *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, pages 1–2. ACM, 2012.

[3] J. Han, X. Xie, and W. Woo. Context-based microblog browsing for mobile users. *Journal of Ambient Intelligence and Smart Environments*, 5(1):89–104, 2013.