Information Sciences 358–359 (2016) 112–128

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Characterizing the interests of social media users: Refinement of a topic model for incorporating heterogeneous media

Jonghyun Han[a,1], Hyunju Lee[b,*]

[a] Defense Agency for Technology and Quality, 420 Dongjin-ro, Jinju, Gyeongsangnam-do 52851, Republic of Korea
[b] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 30 March 2016
Accepted 9 April 2016
Available online 19 April 2016

Keywords:
User interest
Heterogeneous media
Social media
News media
Topic model
Personalization

ABSTRACT

Recent research has focused on extracting personal interest data from social media. Although many methods have been developed, accurately estimating users' interests is often difficult because messages on social media are short and are not classified into any predefined categories. We propose a new method to overcome this problem by incorporating heterogeneous media, such as news. In our method, we first extract explicit features and implicit topics of categories using news media, where implicit topics are determined using a refined topic model. Next, we describe social media messages using these features and topics to estimate users' interests. Compared with several other approaches, our approach provides more accurate estimations of users' interests. We also demonstrate that the accuracy of friend recommendations is increased using the users' interests estimated by our method. Thus, we expect that the proposed approach could be helpful for enhancing the personalization of social media services.

1. Introduction

Online social media services, such as Twitter and Facebook, have recently increased in popularity. Users of these services are able to communicate with each other by posting brief messages that describe their current status or opinions and then share their opinions by following others who have similar interests [11]. The role of social media services differs from that of traditional media, as social media users post more personal-life and pop-culture messages [32]. Hence, social media is an important source for understanding users' interests and providing user-preferred information [1,9,18,31,33].

To obtain information from social media services, users try to find friends who frequently post content that they find attractive. Users may devote time to searching for friends and information related to their interests. Because information searches are usually performed using search functions provided by social media, the quality of the search results depends on the words that are used in the query. For example, if the query is "baseball," a system might recommend other users who have frequently posted the word "baseball" but might not recommend users who have posted words that are related to baseball, such as "homerun" or "MLB." Moreover, the system might not recommend appropriate information because ambiguous words with several different meanings (i.e., homonyms), such as "bank" or "bat," may be used. This issue can

* Corresponding author. Tel.: +82 62 7152213; fax: +82 62 7152204.
E-mail addresses: jhan@dtaq.re.kr (J. Han), hyunjulee@gist.ac.kr (H. Lee).
[1] Tel.: +82-55-7515458, Fax: +82-55-7512204

http://dx.doi.org/10.1016/j.ins.2016.04.020
0020-0255/

lead to insufficient search results. Hence, content should be enriched with information that contains explicitly mentioned terms, as well as information that shows users' implicit interests.

The task of estimating users' interests and recommending relevant information has two main challenges. First, because the content of social media is sparse, retrieving content by only matching queries leads to an information shortage problem. Second, because the data are generated by users, unlike traditional news media, they are not classified into categories. In addition, owing to several issues or novel features specific to social media (i.e., the frequent use of noisy words not related to content, re-tweet functionality and the provision of geographical information), traditional recommendation and information retrieval methods may not be suitable for estimating users' interests in social media services [14,16,27]. To overcome these issues, Hong et al. [14] focused on modeling re-tweet behaviors, and Bernabe-Moreno et al. [3] aimed to quantify the effect of a social media topic on a defined geographical location.

To address the challenges described above, we propose a novel approach in which we estimate social media users' interests by incorporating news media. Social media and news media are heterogeneous in that they have different characteristics. While the content of social media contains unorganized casual mentions of implicit topics, the content of news media is refined by experts and accurately classified into categories. Because categories of news media cover broad topics, themes, and lifestyle interests (e.g., politics, economy, sports, and science), the category information is useful for assigning social media users' interests in a more organized manner. In this paper, we exploit the refined category information of news media to analyze the content of social media. First, we extract two types of feature descriptors that describe categories in news media: implicit topics (e.g., the topic distribution of a category) and explicit features (e.g., the number of times a word appears in a category). Second, these two types of feature descriptors from news media are compared to those obtained from social media in a similar manner. We then express social media users' interests using the categories in news media.

We employ a topic model to identify implicit features of social media and news categories. A topic model discovers latent or implicit topics that occur in documents. However, traditional topic models need to be modified to extract the features of categories from news media or to determine social media users' interests. While topic models, including Latent Dirichlet Allocation (LDA) [4] (the most commonly used topic model), are usually built on the assumption that every term in a document is related to the topic of the document, the content of social media or news media occasionally contains terms not related to the topic of the content. Social media users often mention popular issues despite having little interest in them: "*My timeline is overflowing with the World Cup. Really boring!*" In addition, news articles occasionally contain terms that attract readers' attention regardless of the news topic. An example is a news article about politics with the following sentence: "*A White House petition was created to name Tim Howard, goalkeeper of the US during the 2014 FIFA World Cup, the nation's Defense Secretary*." Although "*World Cup*" and "*Tim Howard*" are not related to politics, news writers use these terms to hold readers' interest. Terms that are mentioned because of currently popular issues but are not related to a real topic are referred to as "issue terms." To focus on terms related to interesting issues and to give less weight to issue terms, we refine a topic model that specializes in handling social media and news media content.

In addition, we propose explicit feature descriptors that employ a hierarchical structure of news categories and terms used in the categories. A hierarchical structure of news categories is useful in addressing a problem caused by commonly used terms among categories. Common terms are often over-weighted compared to specific terms in each category. For instance, "Obama" is an important term for describing the "politics" category, but it is not sufficient to distinguish the subcategories of "politics." To avoid this problem, we apply a different significance to terms according to their level in the hierarchy of categories.

We applied the proposed novel approach to *Twitter* and evaluated its accuracy in estimating social media users' interests. In addition, we recommended friends to social media users based on the estimated users' interests. Our contributions can be summarized as follows:

- We propose an approach that estimates social media users' interests more accurately than the existing approaches that exploit traditional information retrieval methods such as bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF), or LDA [4].
- We refine a topic model that specializes in handling social media and news media. Specifically, the model handles issue terms that are not generated by topics and discovers implicit features of categories.
- We implement an application that recommends other users who have similar interests using real social media and evaluate the results of the application to verify the usefulness of our proposed approach. Our evaluation indicates that utilizing heterogeneous media is useful for providing friend recommendations.

The rest of the paper is organized as follows: Section 2 provides the background for our approach. Our proposed approach to estimating social media users' interests is described in Section 3. In Section 4, we present the results of our evaluation. Our conclusion and suggestions for further research are provided in Section 5.

2. Related works

2.1. Recommendation systems for social media

Recently, many researchers have investigated personalization systems because of the "information overload" caused by the tremendous quantity of available information [2]. In particular, much of the literature has focused on recommendation