Interest Classification of Twitter Users using Wikipedia

Kwan Hui Lim and Amitava Datta
School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
kwanhui@graduate.uwa.edu.au, amitava.datta@uwa.edu.au

ABSTRACT

We present a framework for (automatically) classifying the relative interests of Twitter users using information from Wikipedia. Our proposed framework first uses Wikipedia to automatically classify a user's celebrity followings into various interest categories, followed by determining the relative interests of the user with a weighting compared to his/her other interests. Our preliminary evaluation on Twitter shows that this framework is able to correctly classify users' interests and that these users frequently converse about topics that reflect both their (detected) interest and a related real-life event.

Categories and Subject Descriptors: J.4 [Computer Applications]: Social and behavioral sciences

General Terms: Theory

Keywords: Twitter, Wikipedia, Social Networks, User Interest

1. INTRODUCTION

Many online social networking sites aim to recommend items and content to users based on their indicated interests. However, many users utilize such sites mainly for networking with friends and may not explicitly indicate their interests. Furthermore, these users may be interested in multiple categories with a varying level of interest in each of them. Thus, we propose a framework for classifying the relative interests of these users, with an automated component that utilizes Wikipedia to translate his/her celebrity followings into various interest categories. In addition, our framework classifies the interests of a user with a relative weighting, considering each interest with respect to his/her other interests. Our main contributions include automating the interest classification of celebrities using Wikipedia, introducing a relative weighting to user interest and performing a preliminary evaluation on Twitter.

Various authors have proposed methods to profile user interests based on the messages (tweets) posted by these users [3, 4]. On the other hand, our proposed approach uses topological links (celebrity followings) of users rather than their tweets, thus allowing us to profile (passive) users who may not tweet frequently. Our proposed approach also differs from earlier work in [1, 2] that detect entire communities of users with the same interest as we profile the relative interest of individual users (with a weighting for each interest).

WikiSym '13, Aug 5–7, 2013, Hong Kong, China.
ACM 978-1-4503-1852-5/13/08.
http://dx.doi.org/10.1145/2491055.2491078.

2. OVERALL FRAMEWORK

Our proposed framework is mainly concerned with classifying the interests of a particular user, based on the celebrities that he/she follows and the interest categories these celebrities represent. For a user i, we first proceed to retrieve the set of celebrity users[1] that user i follows, denoted C_i. Thereafter, we classify the interest categories of each celebrity $c \in C_i$ using information on Wikipedia, particularly the "occupation" field and textual write-up on that celebrity's article page.

Specifically, we automate this celebrity interest classification component using a library of 15 different interest categories and a set of keywords associated with each interest category. A celebrity user can belong to one or more of the 15 interest categories if his/her "occupations" field comprises keywords that belong to different interest categories (as listed in our library). If a celebrity does not have an "occupation" field, his/her interest category can also be determined from the first paragraph of his/her article page (e.g. "...is an American professional *basketball player*"). While we demonstrate this keyword-interest mapping with these 15 categories, we can also classify celebrities into other interest categories by building a set of keywords associated with these other interest categories.

Given that a user i follows C_i celebrities, let $|N_{i,I_j}|$ denote the number of celebrities $c \in C_i$ that belong to interest category I_j, for $1 \leqslant j \leqslant 15$ (i.e. the 15 interest categories in our library for the keyword-interest mapping). We now define the relative interest of user i in an interest category I_k as:

$$R_{i,I_k} = \frac{|N_{i,I_k}|}{\sum_{j=1}^{15} |N_{i,I_j}|} \quad (1)$$

In short, R_{i,I_k} is a holistic measure of the interest of user i in interest category I_k, relative to his/her interests in other interest categories. This measurement builds upon the intuitive knowledge that a person who has only one interest is likely to focus more on that interest, compared to another person who has his/her attention spread across multiple interests. Thus, a R_{i,I_k} value closer to 1 indicates that user i has a high interest in category I_k, while a R_{i,I_k} value closer to 0 indicates otherwise.

3. VERIFICATION OF CELEBRITY INTEREST CLASSIFICATION

As one key component of our proposed framework is the correct and automatic classification of celebrities into their representative

[1]While we define celebrities as users with more than 10k followers, this threshold can be adjusted (downwards) to cater for a niche interest category such as Indie Music, English Chess, etc (where its representative celebrities are likely to have less followers).

Table 1: Frequency of Interest Categories (1,000 Celebrities)

Interest	Freq.	Interest	Freq.	Interest	Freq.
Film & TV	415	Media	91	Politics	25
Music	331	Fashion	88	Blogging	20
Publishing	112	Business	75	Charity	17
Sports	105	Hosting	47	Food	3
Comedy	95	Internet	38	Religion	2
				Unclassified	161

interest categories, we first evaluate the performance of this automatic celebrity interest classification component. Our evaluation dataset comprises 1,000 Twitter celebrities who are selected based on their high number of followers. Thereafter, we used the real name of these 1,000 celebrities as input for the automatic celebrity interest classification component. This component then searches Wikipedia for the corresponding article on each celebrity and analyzes keywords used in the "occupation" field or textual write-up of this article. Finally, it compares the detected keywords to our library of keyword-interest mapping and determines the interest categories that each celebrity represents. A celebrity could also represent multiple interest categories such as how a popular "The X Factor" contestant might represent both the Music and Film & TV interest categories.

Our results indicate that this component is able to automatically classify up to 83.9% of all celebrities into their respective interest categories as shown in Table 1. The remaining celebrities could not be automatically classified due to a lack of Wikipedia articles or ambiguous celebrity names (e.g. Pitbull). The lack of Wikipedia articles is due to celebrities who are solely Twitter celebrities but not real-life celebrities (e.g. actors/actress or singers). As such, these Twitter celebrities have a large number of followers (due to their interesting posts) but otherwise do not have a corresponding Wikipedia article (due to them not being real-life celebrities).

In addition, for the 83.9% of celebrities who were successfully classified (automatically), we also manually verified their classified interest categories against their respective Wikipedia article and found little discrepancies with the available information on their respective Wikipedia article. Thus, these results show that the celebrity interest classification component is able to correctly and automatically classify a large majority of celebrities into their respective interest categories.

4. DETERMINING USER INTEREST

After verifying the celebrity interest classification component, we now conduct a preliminary evaluation of the user (relative) interest classification component (represented by Equation 1). For this preliminary evaluation, we chose 172,400 random Twitter users and collected their tweets and follower/following links using the Twitter API, from Nov to Dec 2012.

As described in Section 2, we now determine the celebrities followed by these users and the interest categories represented by these celebrities. Next, we calculate the R_{i,I_k} values (i.e. relative interest) of these users as denoted by Equation 1. In particular, we compare between two groups of users: one group comprising only users with $R_{i,I_k} = 1$ and the other with $R_{i,I_k} \neq 1$, where the interest category I_k is Music. Given the challenges in establishing a ground truth, we best approximate interest (conversational) topics for these two groups based on the Twitter #hashtags used.

Fig. 1 shows a word-cloud of #hashtags that are frequently used by the group with $R_{i,I_k} = 1$. These #hashtags give an indication of the topics of interest to this group and we observe that the two most popular #hashtags are #votecece and #warriorsareproudofcece. These two #hashtags were used to show support and garner votes for a contestant (Cece Frey) on Season 2 of "The X Factor".

Figure 1: Word Cloud of users with $R_{i,I_k} = 1$

Similarly, there are other #hashtags (e.g. #xfactor and #bringjillbackonxfactor) relating to the same TV programme that was showing at the time of our data collection (Nov to Dec 12). Also, another popular #hashtag is #habitamesiempre which is the name of a music album that was released in Nov 12 (and Music is also closely related to the Film & TV interest category).

On the other hand, the group with $R_{i,I_k} \neq 1$ do not display such topical trends compared to the group with $R_{i,I_k} = 1$. The group with $R_{i,I_k} \neq 1$ uses #hashtags that are very diverse and have little or no relation to the Film & TV interest category. These contrasting results for the two groups show that our proposed framework is able to accurately identify user interests based on our preliminary evaluation on Twitter. More importantly, these preliminary results show that any advertising or marketing campaign is likely to command more attention by selecting users with full interest in a category (i.e. $R_{i,I_k} = 1$) as their target audience.

5. CONCLUSION

We presented a framework for classifying the (relative) interests of a Twitter user based on the celebrities that he/she follows. Our proposed framework uses Wikipedia to automatically classify these celebrities into various interest categories (using a keyword-interest mapping). Thereafter, we determine the relative interests of the user (with a weighting) in relation to his/her interest in other categories. In our preliminary evaluation, we have shown that users with a full interest ($R_{i,I_k} = 1$) in the Film & TV category communicate more frequently about this interest (and a corresponding real-life event, "The X Factor" TV programme) compared to users with a divided interest ($R_{i,I_k} \neq 1$) in Film & TV. For future work, we intend to further enhance our framework by considering the interests of the user's friends (social influence), and also perform a comprehensive evaluation on other online social networks.

Acknowledgments: Kwan Hui Lim was supported by the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) under the International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship.

6. REFERENCES

[1] K. H. Lim and A. Datta. Finding Twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd International Workshop on Modeling Social Media*, pages 25–32, Jun 2012.

[2] K. H. Lim and A. Datta. Tweets beget propinquity: Detecting highly interactive communities on twitter using tweeting links. In *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 214–221, Dec 2012.

[3] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, pages 73–80, Oct 2010.

[4] P. Siehndel and R. Kawase. Twikime! - User profiles that make sense. In *Proceedings of the 11th International Semantic Web Conference (Posters and Demos)*, Nov 2012.