Expert Systems with Applications 40 (2013) 1274–1282

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Representative reviewers for Internet social media

Sang-Min Choi, Yo-Sub Han *

Department of Computer Science, Yonsei University, Seoul 120-749, Republic of Korea

ARTICLE INFO

Keywords:
Social-network
Social-media
Influential users
Representative reviewers

ABSTRACT

Our many various relationships with persons from home, work and school give rise to our social networks. In a social network, people receive, provide, and pass a great deal of information. In this process, we often observe that certain individuals have especially strong influences on others. We call these highly influential people opinion leaders. Since the late 20th century, the number of Internet users has increased rapidly, and a huge number of people now interact with each other in online social networks. In this way, the Web community has become similar to real-world society. Internet users receive information not only from the mass media, but also from opinion leaders. For example, online articles posted by influential bloggers are often used as marketing tools or political advertisements, due to their huge influence on other users. Therefore, it is important and useful to identify the influential users in an online society. We thus propose a simple yet reliable algorithm that identifies opinion leaders in a cyber social network. In this paper, we first describe our algorithm for identifying influential users in an online society. We then demonstrate the validity of the selection of representative reviewers using the Yahoo! music and GroupLens movie databases and performing 10-fold cross-validation and z-tests.

1. Introduction

The mass media supplies a great deal of information, which affects us in various ways. However, people often receive information not from the mass media directly, but through the opinions of individuals called opinion leaders. In fact, it seems that opinion leaders have more influence on people than the mass media does (Katz & Lazarsfeld, 1955). While the public does not generally accept information from the mass media uncritically, they do tend to easily accept information from opinion leaders. This suggests that opinion leaders are representatives of their social networks. Since the late 20th century, the number of Internet users has noticeably increased. Especially in the Web 2.0 era, the number of users and the amount of information increased substantially due to the various types of user participation. For instance, many use Yahoo! Answers[1] to post questions and seek answers from other users, and many use Web services such as Google[2] or Wikipedia[3] to search for information available on the Web. These Web-based interactions create a platform where people meet each other and share stories and information. However, not all information on the Web is reliable, and users often encounter spam and misinformation; therefore, people are more likely to trust information from known, trustworthy sites or users. This is similar to behavior in real-world society, in which most people trust information from opinion leaders more than information from the mass media. Recently, as the Web and user activities have grown, online society has come to resemble real-world society in many ways. Studies that have modeled user participation in social networks (Durugbo, 2012), reconciled users opinions on Web applications (Jung, 2012a, 2012b), and analyzed user interactions in various social activities has shown that the Web has characteristics similar to those in real society. On the Web, we can easily find opinion leaders in blogs and social network applications. Their influential opinions appear as articles, posts, or online content, and they are frequently used for marketing tools or political advertisements. Therefore, the current work to create an algorithm that identifies influential users on the Web is useful and meaningful. In Section 2 of this paper, we discuss previous approaches to identifying opinion leaders in Web applications. Then in Section 3, we describe our algorithm that identifies representative reviewers in Internet social media. In Section 4, we provide verification of our approach through test results that show the validity of the algorithms choices for representative reviewers, and we test reliability using 10-fold cross-validation and z-tests and using the Yahoo! music[4] and GroupLens movie databases.[5] Finally, we provide conclusions and indicate future directions for this research in Section 5.

* Corresponding author. Tel.: +82 2 2123 5725; fax: +82 2 365 2579.
E-mail addresses: jerassi@cs.yonsei.ac.kr (S.-M. Choi), emmous@cs.yonsei.ac.kr (Y.-S. Han).

[1] http://answers.yahoo.com/.
[2] http://www.google.com/.
[3] http://en.wikipedia.org/.
[4] http://webscope.sandbox.yahoo.com/.
[5] http://grouplens.org/.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.063

2. Related study

In this section, we briefly describe previous research on methods that find opinion leaders in Web applications. There are two main approaches: one based on social relations and one based on contents metadata. For example, Kwak, Lee, Park, and Moon (2010) examined opinion leaders on the social-relations Web application, Twitter,[6] which composes a social network by allowing users to connect and communicate with each other using such functions as follow, reply, retweet, and post. They found that the characteristics of users with high numbers of followers are different from users with a high number of retweets, with followers reacting more sensitively to tweets by users with many retweets. In another study based on social relations, Han, Kim, and Cha (2012) described a social network based on video contents and user activities such as subscription, uploading and favorite. They used a modified PageRank algorithm to calculate user reputation in this contents-based social network, finding that it was closely related to subscriptions and the number of uploads. They proposed an algorithm that composed the social-network from users of the video contents and derived user reputation based on uploading and subscription. Another approach is based on contents metadata, such as the algorithm proposed by Agarwal, Liu, Tang, and Yu (2008) that identifies an influential user on blog sites. An important function of blogs is to post various media contents and to tag relevant information from other blogs. The influence of blogs on users in the Web can be strong; for example, the preferences of opinion leader bloggers can affect the purchases made by their visitors, making it profitable to advertise on influential blogs. Various studies have examined the essential issues of identifying influential bloggers, evaluated the effects of various collectible statistics from a blog site to determine blog-post influence, developed unique experiments using Digg,[7] and conducted experiments using whole histories of blog posts. The research to identify influential bloggers has classified the characteristics of bloggers into active, inactive, influential, and non-influential bloggers based on intuition. Active bloggers are those who often create posts, and influential bloggers are those whose posts influence others as determined by social gestures such as comments, incoming links, outgoing links, and lengths of posts. This study clearly demonstrated the existence of influential bloggers, and how influential bloggers relate to each other and to other common visitors. In another study that identified influential users based on contents metadata, Cha, Lee, Han, and Kim (2009) proposed a method to evaluate user reputation on questioning sites such as Yahoo! answers. They collected the n-gram from a title or from a question and its answers, and they calculated similarity using their proposed equation. They then determined user reputation using a modified PageRank algorithm, with scores given to links between questions and answers.

3. Our approach

3.1. Identifying representative reviewers

We propose an algorithm that identifies the representative reviewers of an evaluative group for Internet social media. By representative reviewers, we mean the users who have high representativeness for media contents as rated by many users. The representative reviewers can express evaluations of the other reviewers. Thus, they are similar to influential users in real society. To identify representative reviewers, we use Eq. (1).

$$U_s = \frac{\sum_{i \in A} |(R_S(i) - R_\mu(i)|}{|A|}, \quad (1)$$

[6] http://www.twitter.com/.
[7] http://digg.com/.

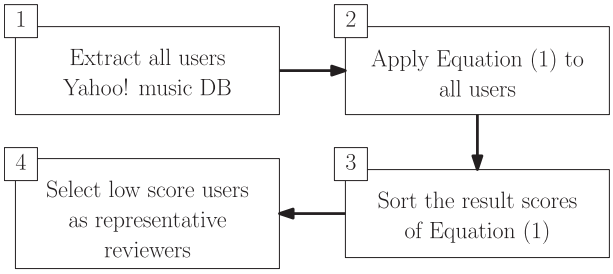

Fig. 1. The procedure of identifying representative users from Yahoo! music dataset.

Table 1
Yahoo! music database.

Attribute	Explanation
UserID	There exist 15,400 users in total, and given by integer number from 1 to 15,400
ItemID	There exist 1,000 songs in total, and given by integer number from 1 to 1,000
Rating	As integer number from 1 to 5 there exist approximately 300,000 ratings

Table 2
GroupLens movie database.

Dataset	Attribute	Explanation
Movie dataset	MovieID, title, genre	There are total of 10,681 movies
User dataset	UserID, gender, age, occupation, zip-code	There are total of 69,898 users
Rating dataset	UserID, movieID, rating, timestamp	There are total of 10,000,054 ratings

Fig. 2. The distribution of scores (U_s) from Eq. (1) in ascending order for the Yahoo! music database.

where A is a set of contents rated by user S, and $|A|$ is the cardinality of A. $R_S(i)$ is the rating of content i by user S, and $R_\mu(i)$ is the average rating of content i by user S. Note that the result of Eq. (1) shows how close each users rating is to the average rating. We select those who have low scores from Eq. (1) as representative reviewers.

Fig. 1 shows the procedure for identifying representative reviewers in the contents raters group. First, we extract all users who evaluate media contents. Second, we apply Eq. (1) to all these extracted users. Third, we sort the result scores of Eq. (1) in ascending order. Finally, we select the low-score users as the representative reviewers.