Knowledge-Based Systems 51 (2013) 35–47

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Twitter user profiling based on text and community mining for market analysis

Kazushi Ikeda [a,*], Gen Hattori [a], Chihiro Ono [a], Hideki Asoh [b], Teruo Higashino [c]

[a] KDDI R&D Laboratories, Inc., 2-1-15, Ohara, Fujimino, Saitama 356-8502, Japan
[b] National Institute of Advanced Industrial Science and Technology, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
[c] Graduate School of Information Science and Technology, Osaka University Yamadaoka, 1-5, Suita-shi, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Received 22 August 2012
Received in revised form 27 June 2013
Accepted 29 June 2013
Available online 12 July 2013

Keywords:
Web mining
Market analysis
User profiling
Twitter
Text analysis
Community analysis
Machine learning

ABSTRACT

This paper proposes demographic estimation algorithms for profiling Twitter users, based on their tweets and community relationships. Many people post their opinions via social media services such as Twitter. This huge volume of opinions, expressed in real time, has great appeal as a novel marketing application. When automatically extracting these opinions, it is desirable to be able to discriminate discrimination based on user demographics, because the ratio of positive and negative opinions differs depending on demographics such as age, gender, and residence area, all of which are essential for market analysis. In this paper, we propose a hybrid text-based and community-based method for the demographic estimation of Twitter users, where these demographics are estimated by tracking the tweet history and clustering of followers/followees. Our experimental results from 100,000 Twitter users show that the proposed hybrid method improves the accuracy of the text-based method. The proposed method is applicable to various user demographics and is suitable even for users who only tweet infrequently.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Recently, due to the widespread popularity of the Internet, many people state their opinions via social media services. In particular, Twitter [1] is a suitable platform for real-time, casual communication. Many Twitter users post opinions about products, services, and TV programs. It is essential for companies to make efforts to improve their products and services based on their customers' requirements. As a means of using user opinions for marketing, reputation analysis technologies have recently attracted a great deal of attention [2,3]. Compared to previous marketing approaches based on questionnaire surveys, online opinion analysis has many advantages, including real-time feedback, low cost, and high volume. User demographics such as age, gender, and residence area are also essential for marketing analysis, since opinions vary with user demographics. For example, functions of mobile phones that are popular among young people are often found awkward to use by the elderly. Since most Twitter users do not state their demographic information, it has been impossible to extract opinions for individual user demographic segments (such as teens, twenties, or thirties). Several text-based approaches have been proposed to extract user demographic information [4–6]. However, only few proposals for large-scale and practical marketing analysis applications to perform demographic estimation exist due to difficulties in improving the effectiveness and accuracy to a level sufficient for practical use. Considering practical use, we realized that a general approach is required for estimating wide varieties of demographics such as age, gender, area and other categories. An estimation method targeting users with few tweets such as followers of corporate accounts is also important.

To solve these problems, we propose a hybrid of a text-based method and a community-based method for the demographic estimation of Twitter users. The text-based method estimates the demographics of users whose tweets contain sufficient text features. For all other users, the community-based method analyzes the followers/followees whose tweets contain plentiful text features. The hybrid method covers almost all users by making the most of the Twitter platform, including both tweets as text information and followers/followees as community information. In the text-based method, characteristic terms used by each demographic segment are automatically detected based on linguistic and statistical analysis by tracking the content of users' tweet histories. For example, users whose tweets often include terms such as "school," "classroom," and "examination" are presumed to be teens and students. In the community-based method, demographic

* Corresponding author.
E-mail addresses: kz-ikeda@kddilabs.jp (K. Ikeda), gen@kddilabs.jp (G. Hattori), ono@kddilabs.jp (C. Ono), h.asoh@aist.go.jp (H. Asoh), higashino@ist.osaka-u.ac.jp (T. Higashino).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.06.020

information is estimated from the follower/followee relations of the target user. In the proposed method, characteristic biases in the demographic segments of users are detected from the community groups constructed by clustering their followers and followees. A user can have several community groups, such as local friends, co-workers and hobby groups, where the members of each group have something in common such as age, gender and regional area.

Social opinions and demographic information are extremely attractive to businesses. For instance, product planners need to understand user requirements, customer support and service management departments need to monitor customer responses, advertising agencies want to deliver persuasive advertisements to target audiences, and broadcast TV directors need real-time feedback from the audience. In this paper, we focus on Japanese Twitter users. However, the algorithms of the proposed text-based and community-based methods are applicable to any language.

The rest of the paper is structured as follows. We outline related work in Section 2. We describe the proposed text-based method, community-based method and hybrid method for demographic estimation in Section 3 and the results of performance evaluations in Section 4, respectively. We conclude this paper in Section 5.

2. Related work

Extracting author information from the Web has been attempted for a long time. Table 1 summarizes the previous works. An extraction method for author information from Web sources is proposed for the purpose of judging whether the information is trustworthy [7]. Koppel et al. classify three author attribution problems [8]: (1) the profiling problem, where the challenge is to provide as much demographic or psychological information as possible about the author [4–6]; (2) the needle-in-a-haystack problem, where there are many thousands of candidates for each of whom we might have a very limited writing sample [9]; and (3) the verification problem, where the challenge is to determine whether the target is the author or not [10]. The problem that we tackle in this paper is related to (1).

Common approaches to author profile estimation from documents use the volume of each term in the document for classification. Argamon et al. estimate the authors' age, gender, native language, and personality from blogs and essays written by university students [4]. Estival et al. estimate age, gender, nationality, education level, and native language from English e-mails [5]. Pham et al. estimate age, gender, and area from Vietnamese blogs [6]. However, in these previous studies, the evaluations are only on small platforms, such as blogs, essays, or e-mails. With practicality in mind, we propose a profile estimation method on Twitter, which is one of the largest, most popular, and most internationally accepted platform among social media.

There are challenges associated with the author attribution problem of (2) and (3) on the Twitter platform [9,10]. Layton et al. show that the important threshold is 120 tweets per user, at which point adding more tweets per user gives a small but non-significant increase in accuracy for the author attribution problem [11]. Silva et al. show that markers include highly personal and idiosyncratic editing options, such as emoticons, interjections, and punctuation, which are often seen in casual SNS (Social Networking Services) such as Twitter [12]. In these studies, only text information is used, which is considered to limit accuracy. We propose a hybrid method comprising a text-based method and a community-based method, which enhances the accuracy.

Follower and followee relationships are regarded as directed links between two users. There has been some research reporting link-based document classification methods, such as for scientific papers based on co-citations [13] and for Web pages based on their hyperlinks [14]. Hybrid methods composed of text-based methods and link-based methods are reported to improve the accuracy of Web page classification [15–18]. Calado et al. show that classification methods based on co-citation are effective [15]. Qi and Davison show that the content and topic information of neighboring documents increases classification accuracy [16]. Zhang et al. propose an optimization method with text-based and graph structures for categorization of Web pages [17]. These contributions are helpful for improving the performance of text-based methods. The methods are designed on the assumption that a Web page belongs to only one category at a time. In the demographic estimation problem, however, a user has multiple demographics such as age, gender and area. In that case, existing clustering algorithms for Web pages are not simply applicable to the problem. Therefore, we propose a demographic estimation method targeting user communities.

We have previously proposed a text-based demographic estimation method for Twitter users and its application to broadcast TV programs [19]. In this previous work, the demographic estimation method is limited in terms of the minimum functions required for consumer usage. Only the basic demographic categories such as age, gender and area of residence are estimated. Estimation for users with few tweets is outside of its scope. Going beyond the previous work, we have conducted inquiry surveys in five depart-

Table 1
Summary of previous works related to extraction of author information.

Methods	Information source	Problems (target profiles)	Algorithms
Profile estimation			
Argamon et al. [4]	English blog and essay	Age, gender, native language and personality	Text-based classification
Estival et al. [5]	English e-mail	Age, gender, nationality, education level and native language	Text-based classification
Pham et al. [6]	Vietnams blog	Age, gender and area	Text-based classification
Ikeda et al. [19] (authors' previous work)	Twitter (Japanese tweet)	Age, gender and area	Text-based classification
Proposed method (this paper)	Twitter (Japanese tweet)	Age, gender, area, hobby, occupation and marital status	Hybrid of text-based and community-based
Other related work			
Kato et al. [7]	Web page	Increase credibility of information	Extract authors information
Abbasi and Chen [9]	Email, Web pages and chat	Find an author from thousands of candidates	Text-based information retrieval
Koppel et al. [10]	Literatures	Answer a given target text is or is not written by a given author.	Text-based classification
Layton et al. [11]	Twitter	Survey of the number of tweets required for the author attribution problems	Text-based classification
Silva et al. [12]	Twitter	Improve performance of author attribution problems	Text-based classification including markers

Download English Version:

https://daneshyari.com/en/article/402733

Download Persian Version:

https://daneshyari.com/article/402733

Daneshyari.com