Expert Systems with Applications 42 (2015) 5256–5263

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Am I typing fresh tweets: Detecting up-to-dateness and worth of categorical information in microblogs

Mustafa Özgür Cingiz, Banu Diri, Göksel Biricik *

Computer Engineering Department, Yildiz Technical University, 34220 Esenler, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online 7 March 2015

Keywords:
Microblog categorization
Short text classification
Social media
Twitter

ABSTRACT

Microblogs are one of the most popular social network areas where users share their opinions, daily activities, interests or other user content. As microblogs generally pose the user's interests, the field of interests can be extracted by using the presented content. In this study, we group microblog users as normal or bot depending on their supplied content and evaluate the user groups with respect to how well they reflect their categories with fresh entries, essentially by using content mining. Traditional content mining studies do not evaluate whether the supplied user entries are up-to-date or not. Unlike similar studies, we check up-to-dateness of users' content by simultaneously retrieving user entries and RSS news feeds. If a term of user content is absent in the feature set that is formed by RSS news feeds, it is not regarded as a feature to check the freshness of the content. For each user group, we divide users into predefined categories and inspect how well the group users post relevant entries while checking the up-to-dateness of their content. Our experimental results prove that bot users always post fresher and category-relevant entries. Finally, we visualize the categorization performances of each user group's entries with Cobweb. The Cobweb presentation unveils the miscategorization tendencies of the user groups.

1. Introduction

After Web 2.0 concept emerged, people are not simple content readers anymore since they can also contribute content as writers. Web 2.0 introduces concepts such as social networks, blogs and microblogs to users. Microblogs are simply defined as short text messages which the writers share immediate information or comments through, on virtually any category. Microbloggers share their opinions, feelings, images, favorite videos and other user's contributions throughout their microblog content (Efron, 2011). Twitter is one of the most popular microblog applications because of its easy sign-up process, ease-of-use and mobile accessibility.

Knowledge discovery in microblogs, especially in Twitter is a trending research area in the recent years (Rao, Yarowsky, Shreevats, & Gupta, 2010). Besides the textual content, other user-specific data like retweet count, user-interaction records, followed users and follower graphs are frequently used. Together or separately, these two kinds of data drive applications in categorical user classification, topic analysis, opinion mining, influential leader discovery, recommendation systems and interestingness discovery. Although these research areas have different names, both share a common objective and use similar features. For example, by evaluating the content of a successfully categorized user, we can either recommend the user for that category, or label as an influential leader. Although, the studies lack checking up-to-dateness of the supplied content.

In Twitter, users can follow other users according to their field of interest. Followers expect tweeting users to feed relevant tweets for the categories that they are interested in. For example, a football player whose team is supported by the followers is expected to enter content about the team that he plays. This elemental requirement makes categorization crucial, especially based on the tweet contents. This is a challenging task because of two reasons. First, the data size is enormous: About 500 million tweets are posted on a typical day. Second, the nature of the tweet contents are ill-posed: They are noisy, contain many abbreviations and reduced text, due to the 140 character limitation. Yet, there are tools like wefollow.com or the search engine in Twitter that users can find out a microblogger's category information. Furthermore, there are many fashionable categorization studies in the literature like influential leader and popular topic identification but despite their popularity, we see that none of them consider the up-to-dateness of the content. These studies do not guarantee that the influential leaders are really tweeting about the topic they belong to or the content is relevant to the category (Ma et al., 2013; Yang & Rim,

* Corresponding author.
E-mail addresses: mozgur@ce.yildiz.edu.tr (M.Ö. Cingiz), banu@ce.yildiz.edu.tr (B. Diri), goksel@ce.yildiz.edu.tr (G. Biricik).

http://dx.doi.org/10.1016/j.eswa.2015.02.025
0957-4174/

2014). The motivation of this study arise from this problem point. There should be a reliable method to discover whether the user groups that Twitter users follow under a certain category post both topic-relevant and fresh tweets, or not.

This motivation addresses three research problems:

1. Classification of the user's tweets as up-to-date or not.
2. Comparison of the classification results in bot and normal users.
3. Visual observation of the intersections and clutter of the user groups (Benjamin et al., 2014).

Moving ahead from our motivation, in this paper we intend to discover how well and fresh the microbloggers reflect their category. To achieve our intention, we use both normal "tweet writers" and news bots that post automated tweets. We visualize and discuss categorical intersections of the user groups with Cobweb representation.

The presented research problem has several handicaps which makes the solution more demanding. Tweet categorization needs a content-based text-mining approach, but the content length is limited 140 characters. Thus, tweets usually consist of abbreviations and a social-media based jargon that are meaningless to the normal written language. These types of irregular entries decrease the classification success rate because it becomes harder to properly tokenize and represent the content (Clark & Araki, 2011). Elimination of these features leads us to an increase in the classification performance. Thus, a careful inspection and preprocessing of the content is the key to success (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010). Microblog users not only post tweets involving their categories but also chat between each other in daily life. This kind of usage also scrambles the categorical knowledge discovery process (Arnott, Goli, Bradley, Smith, & Wilson, 2014). In order to identify how the tweets match to the up-to-date categorical content, we use concurrent RSS news feeds, as they provide reliable and definitive summary for a given category.

In this study, we divide the microblog users in two groups as normal and bot users. User group identification is another related classification problem as both groups have similar behavior patterns (Bosshart & Schoenhagen, 2013; Edwards, Edwards, Spence, & Shelton, 2014). We assume that users interacting with other users with replying, retweeting or favorite labeling are in the normal users group. On the other hand, non-interacting users are pretended to be the bot users. The up-to-date categorical tweeting performances of these user groups are compared based on the RSS contributed content classifications. Misclassifications can not be represented well with traditional performance metrics. For this reason, we choose Cobweb representation as it effectively visualizes the categorized user contents (Benjamin et al., 2014). We visualize the categorical interferences of 120 normal and bot users in 15 categories using Cobweb. The outline of our method also addresses our contributions. We identify the worth of categorical information and up-to dateness of the content presented by user groups. The contributions of this study can be listed as:

- The up-to-dateness of tweets with respect to their categories are controlled. This is achieved by concurrently collecting both RSS news feeds and writers' tweet contents. Control is done be filtering the shared terms.
- We unfold to what extent the user groups reflect their categories, by using different classifiers. 15 categories from wefollow.com application are selected for the evaluation of 120 users in two groups.
- We visualize the misclassifications using Cobweb representation, which provides a better understanding of error patterns of the user groups in the selected categories.

With the contributions stated above, our method can easily be extended to other study fields like influential leader or popular content discovery. An up-to-date tweeting user under a category has a high potential to be an influential leader. Also, fresh tweets guarantee popular content. User recommendation under selected topics can also be achieved with our approach. These are some of the practical implications of our approach, that are possible to achieve with our contributions.

This paper is organized as follows. In the second section, we briefly review microblog text processing literature. In the third section, we introduce our data sets and their features as our experimental material, along with the preprocessing operations. This section also introduces methods utilized for classification. Then we introduce our approach to locate users whose content is more valuable for the related categories, with its model and processing steps. Concluding, we discuss classification results and visualize them using Cobweb.

2. Related work

The extensive microblog usage provide enormous unprocessed data which draws attention of various working fields. Categorical user classification, topic analysis, opinion mining, influential leader discovery, user or product recommendation systems, popular content discovery are among the most popular studies that utilize the content in social networks. These studies mainly rely on classification techniques and locate users who share similar interests or categorize the topics and the contents. Another streamline is pattern discovery in microblogs, having opinion classification and sentiment analysis as the locomotive study fields.

User classification studies use the microblog content, linguistic features, user profile data, social network user interaction topologies for identifying the categories (Liang et al., 2014). The approach in Pennacchiotti and Popescu (2011) classifies microbloggers that include profile information, tweeting behavior, linguistic content of user messages and social network features in three different sectors. In the first sector, 10,338 democrats and republicans are classified and classification performance scored more than 88.9%. In the second sector, classification of ethnicity performed 65.5% success rate. In the third sector, Starbucks fans are specified with 75.9% performance. A relative study Aksu Degirmencioglu and Uskudarli (2010) extracts word-hashtags, hashtag-user and word-user pairs from tweets in order to discover users' common interest areas. The search system developed in Liang et al. (2014) matches keywords about academics, occupations and companies to classify microblog users for recommendation in these three categories. Rao et al. (2010) uses SVM to classify user attributes by using a rich set of features. The study exhibits that distinctive language usage in Twitter reveals latent user information of age, gender, regional origin and political orientation. In Aslan (2010), news pattern similarity is used to discover microbloggers who broadcast news. In the study, microblogger content and text classification techniques are used for measuring the convenience of users' categories.

Opinion mining and sentiment analysis on microblogs are other popular fields in user categorization, where the linguistic features of user content are excessively allocated in the proposed methods. Ren and Wu (2013) use social context and topical context to predict the positive and negative polarity of user content under certain hashtags, and compare their proposed approach with other collaborative filtering methods. Perez-Tellez, Pinto, Cardiff, and Rosso (2011) investigate impact of tweets on reputation of companies. They disambiguate content of tweets for discovering important knowledge about companies. Another study Aisopos, Papadakis, Tserpes, and Varvarigou (2012) works on sentiment analysis over tweets by extracting textual patterns in content and specifying