Online Social Networks and Media 5 (2018) 23–36

Contents lists available at ScienceDirect

Online Social Networks and Media

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

Hybrid semantic clustering of hashtags

Ali Javed, Byung Suk Lee*

Department of Computer Science, University of Vermont, Burlington, Vermont, USA

ARTICLE INFO

Article history:
Received 19 February 2017
Revised 28 October 2017
Accepted 28 October 2017

Keywords:
Hybrid clustering
Semantic clustering
Hashtag
Social media

ABSTRACT

Clustering hashtags based on their semantics is an important problem with many applications. The uncontrolled usage of hashtags in social media, however, makes the quality of semantics and the frequency of usage vary a lot, and this poses a challenge to the current approaches which capitalize on either the lexical semantics of a hashtag (by using metadata) or the contextual semantics of a hashtag (by using the texts associated with a hashtag). This paper presents a *hybrid* semantic clustering algorithm that uses the complementary strengths of lexical and contextual semantics of a hashtag to produce accurate clusters on a wider range of input data. The hybrid algorithm uses a consensus clustering approach, which finds the consensus between metadata-based sense-level semantic clusters and text-based semantic clusters. A gold standard test shows that the hybrid algorithm outperforms both the text-based algorithm and the metadata-based algorithm for a majority of ground truths tested and that it never underperforms both base algorithms. In addition, a larger-scale performance study, conducted with a focus on disagreements in cluster assignments between algorithms, show that the hybrid algorithm makes the correct cluster assignment in a majority of disagreement cases.

1. Introduction

A hashtag is “a word or phrase that starts with the symbol # and that briefly indicates what a message (such as a tweet) is about” [1]. Chris Messina first proposed to use ‘#’ on Twitter in August 2007, to tag topics of interest [2]. Hashtags are now used in social media for all sorts of reasons – to tell jokes, follow topics, launch campaigns, put advertisements, collect consumer feedback, and much more. #OccupyWallStreet, #ShareaCoke and #NationalFriedChickenDay are just a few examples of many successful hashtag campaigns. McDonald’s created hashtag #Mcdstories to collect consumer feedback.

Since Twitter is the first social media platform that introduced hashtags, it is used as the representative social media in this paper. It is estimated that, as of January 2016, Twitter has about 332 million active monthly users uploading 500 million tweets per day. A tweet is a string up to 140 characters, and most tweets contain one or more hashtags in them.

Clustering is a well-known data mining technique for dividing items into groups (or “clusters”) such that items within the same cluster tend to be more similar to each other than those in different clusters [3]. Clustering is commonly used as a text classification technique [4], and, as asserted by Vicient and Moreno [5], clustering of hashtags is the first step in the classification of tweets given that hashtags are used to index those tweets. Therefore, it can be argued that classification of tweets benefits from accurate clustering of hashtags.

Further, on average 6000 messages are posted per second [6] on Twitter alone, making social media arguably the best source of timely information. In this regard, social media analysts use clusters of hashtags as the basis for more complex tasks [7], such as retrieving relevant tweets [7,8], tweet ranking, sentiment analysis [9], data visualization [10], semantic information retrieval [11], and user characterization. Therefore, the accuracy of hashtag clustering is important to the quality of the information resulting from those tasks.

Hashtag clustering has real world impacts. For instance, it can be used to improve the user engagement in social media activities. Social media websites typically use posts (e.g., tweets) on “home timelines” to increase the level of user engagement. Posts may appear on a user’s home timeline for a number of reasons – because they are shared by the user’s direct contacts, because they are publicly disseminated as popular posts, and because they are advertisements sponsored by commercial entities. Given that a hashtag is a viable representation of the posts, accurate clustering of hashtags can improve the content rendering of those timelines for certain users by introducing posts that are beyond their social

* Corresponding author.
E-mail addresses: ajaved@uvm.edu, alijaved@live.com (A. Javed), bslee@uvm.edu (B.S. Lee).

https://doi.org/10.1016/j.osnem.2017.10.004
2468-6964/

network but relevant to their interests as gauged by the hashtags in their posts. In another instance, the categorization of users, resulting from clustering their posts by hashtag, can help advertisement agencies find new potential customers.

There are two major approaches to clustering hashtags. One approach identifies the *lexical* semantics of hashtags from external resources (i.e., "metadata") independent of the tweet messages themselves [5,12,13]. The other approach does that from the tweet texts (i.e., "data") accompanying hashtags [7,10,11,14–17] by identifying their *contextual* semantics [18].

Performance of the metadata-based approach depends on two factors – metadata quality and hashtag quality. It is out of question that the quality of the metadata has a direct impact on the performance. As importantly, with no syntactic or semantic control over the message content, it is common that hashtags contain errors and abbreviations, thus hampering metadata search quality because of poor quality of the search input.

The metadata-based approaches at the present time are a relatively new area of research that is benefiting from the increasing availability of metadata. This approach has the advantage of being immune to poor linguistic quality of tweet messages that contain hashtags, but has the disadvantage of being sensitive to the quality of metadata or the degree of match between them and hashtags.

There have been more works using the text-based approach [7,10,11,14–17]. In this approach, tweet messages are compared using the *bag-of-words* model [19], and thus the performance depends largely on the amount of text associated with the hashtag. This approach has the advantage of being largely unaffected by poor linguistic quality of hashtag and being able to span across all languages (including slang/informal languages).

It, however, has the disadvantage of working well only on *common* hashtags, as uncommon hashtags do not have enough tweet messages accompanying them. As cited by Tsur et al. [15], 1000 most popular hashtags, which comprise 0.003% of all distinct hashtags, cover about 43% of over 417 million tweets in their corpus – this puts the performance of the bag-of-words approach in question for the remaining 99.997% of hashtags.

Thus, the current approaches to semantic hashtag clustering do not possess the *versatility* needed to produce accurate clusters under varying circumstances, that is to say, all common or rare English language hashtags with varying semantic quality. The sources of hashtag semantics used in the current approaches are orthogonal to each other and their performances are complementary to each other. Hence, this paper aims to combine the two approaches into a *hybrid* approach. The aim is that the hybrid algorithm produces accurate results on a wider range of input data. Such a versatile algorithm unburdens the user from having to decide which algorithm to use for accurate results when there is no ground truth available or when the tweet dataset is so arbitrary that it is not clear which approach is better.

Thus, this paper addresses the problem of clustering hashtags based on two kinds semantics – lexical from metadata and contextual from texts. For this purpose, two base algorithms, each specializing in the respective semantic sources, are utilized and the hybrid semantics combining the two sources are realized by building a consensus from the results of the two base algorithms. To the best of our knowledge, this paper is the first one that addresses combing two distinct semantic sources, namely "lexical" and "contextual", to identify the semantics of hashtags for a certain task, e.g., clustering.

Specifically, we design a hybrid semantic clustering algorithm using two base algorithms, each representing one of the two approaches. The first base is the metadata-based semantic hashtag clustering algorithm introduced in our prior work [12,13] enhanced from the original algorithm by Vicient and Moreno [5]. The second base is the text-based semantic hashtag clustering algorithm adapted from the algorithm proposed by Tsur et al. [15,16] and Muntean et al. [7], which uses the bag-of-words model. Output clusters of these two base algorithms are input to the hybrid algorithm. This hybrid algorithm is based on the concept of *consensus clustering*, as a mere intersection of the two outputs would be too restrictive and not scalable (if more base algorithms were to be added later).

Our hybrid clustering is unique in that what it combines are the two distinct, yet complementary sources of semantics (i.e., lexical and contextual) on the same clustering method (e.g., hierarchical clustering), while other existing body of work on hybrid clustering (e.g., [20–23]) combine two distinct clustering methods. Additionally, no existing hashtag clustering algorithm utilizes multiple distinct sources of semantics to produce more accurate results on a wider range of data, thus validating the complementary nature of semantics used.

Our hybrid clustering algorithm was evaluated using two different experiments – a gold standard test and a "pairwise disagreement" test. The gold standard test showed that two, among the three (i.e., hybrid and the two base) algorithms, the hybrid algorithm achieved the highest accuracy for 57% of ground truth data sets and the second highest accuracy for the remainder (i.e., 43%) of them, and in this case the gap with the better one was marginal (i.e., 10% to 17% in "weighted average pairwise maximum f-score"). The pairwise disagreement test was done with a focus on the instances of disagreement occurring in clustering decision between the hybrid and the base algorithms, where a decision was made for each pair of hashtags whether to cluster them together or to separate them. The result showed that the hybrid clustering made the right clustering decision more than 90% of the time when there were disagreements. In addition, we present anecdotal examples from the clustering results to demonstrate the merit of the hybrid approach. Overall, the experiment results confirm that the performance of the hybrid approach is more versatile than either of the two underlying algorithms individually in various environments, thus demonstrating how these two different algorithms complement each other to hold up the performance together as a hybrid even when one algorithm performs poorly.

All source codes and datasets, including the gold standards, are available from Github at https://github.com/ali-javed/hybrid_semantic.

The remainder of the paper is organized as follows. Section 2 provides some background knowledge. Section 3 discusses related work. Section 4 discusses the base algorithms used in the design of the hybrid algorithm. Section 5 presents the details of the hybrid algorithm and its evaluation against the two base algorithms. Section 6 summarizes the paper and suggests future work.

2. Background

This section provides some background knowledge needed for the readers to understand this paper.

2.1. WordNet – synset hierarchy and similarity measure

WordNet is a free and publicly available lexical database of English language [24]. It groups words into sets of synonyms called synsets. Each word in WordNet must point to at least one synset, and each synset must point to at least one word. Hence, there is a many-to-many relationship between synsets and words. Synsets in WordNet are interlinked by their semantics and lexical relationships, which results in a network of meaningful related words and their senses.

Table 1 shows an example synset. The synset contains four different senses – e.g., "desert" meaning "arid land with little or