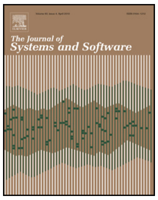
The Journal of Systems and Software 94 (2014) 16–29

Contents lists available at ScienceDirect

The Journal of Systems and Software

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



Twitter data analysis by means of Strong Flipping Generalized Itemsets

Luca Cagliero[*], Tania Cerquitelli, Paolo Garza, Luigi Grimaudo

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form 28 January 2014
Accepted 19 March 2014
Available online 28 March 2014

Keywords:
Social network analysis and mining
Data mining and knowledge discovery
Generalized itemset mining

ABSTRACT

Twitter data has recently been considered to perform a large variety of advanced analysis. Analysis of Twitter data imposes new challenges because the data distribution is intrinsically sparse, due to a large number of messages post every day by using a wide vocabulary. Aimed at addressing this issue, generalized itemsets – sets of items at different abstraction levels – can be effectively mined and used to discover interesting multiple-level correlations among data supplied with taxonomies. Each generalized itemset is characterized by a correlation type (positive, negative, or null) according to the strength of the correlation among its items.

This paper presents a novel data mining approach to supporting different and interesting targeted analysis – topic trend analysis, context-aware service profiling – by analyzing Twitter posts. We aim at discovering contrasting situations by means of generalized itemsets. Specifically, we focus on comparing itemsets discovered at different abstraction levels and we select large subsets of specific (descendant) itemsets that show correlation type changes with respect to their common ancestor. To this aim, a novel kind of pattern, namely the Strong Flipping Generalized Itemset (SFGI), is extracted from Twitter messages and contextual information supplied with taxonomy hierarchies. Each SFGI consists of a frequent generalized itemset X and the set of its descendants showing a correlation type change with respect to X.

Experiments performed on both real and synthetic datasets demonstrate the effectiveness of the proposed approach in discovering interesting and hidden knowledge from Twitter data.

1. Introduction

In recent years, social networks and online communities have become a powerful source of knowledge. Social network sites, such as Twitter and Facebook, are accessed by millions of people every day and their users continuously publish multimedia resources, posts, and blogs.

Since actions undertaken by Web users reflect their habits, personal interests, and professional skills, a particular attention has been paid to the analysis of data acquired from Twitter. Although a large body of research addresses social network data mining (Cagliero and Fiori, 2013; Cagliero et al., 2013; Cheong and Lee, 2009; Kasneci et al., 2009; Li et al., 2010, 2008; Mathioudakis and Koudas, 2010; Yin et al., 2009), the potential business impact of mining social data is still largely unexplored. Service providers (e.g., TV channels, radio stations) may explore and analyze Twitter posts to improve service provision according to the knowledge hidden in Twitter data. From a business point of view, it is worth profiling Twitter user trends and message topics to plan targeted promotions or to identify exceptional situations. For example, let us consider a music tour organizer. To plan a singer tour in Italy, it is important to known in which Italian cities and in which time periods people are most likely to attend concerts. The analysis of the Twitter messages posted by Italian users about a given singer may support organizers in planning tours and advertising sessions. For example, they can promote album or songs to specific user segments. Innovative analytics solutions are needed to effectively and efficiently support service profiling and topic trend analysis as well as to discover exceptional situations from large social data collections. However, Twitter data is intrinsically sparse because posts range over many different topics and use a wide vocabulary. To address this issue, a promising research direction is to exploit semantics-based models (e.g., ontologies, taxonomies) to drive the data mining process and to discover interesting correlations among Twitter data at different abstraction levels.

[*] Corresponding author. Tel.: +39 011 090 7084; fax: +39 011 090 7099.
E-mail addresses: luca.cagliero@polito.it, luca.cagliero84@gmail.com (L. Cagliero), tania.cerquitelli@polito.it (T. Cerquitelli), paolo.garza@polito.it (P. Garza), luigi.grimaudo@polito.it (L. Grimaudo).

http://dx.doi.org/10.1016/j.jss.2014.03.060
0164-1212/© 2014 Elsevier Inc. All rights reserved.

Generalized itemset mining (Srikant and Agrawal, 1995) is an exploratory data mining technique that allows us to discover multiple-level correlations among data supplied with an analyst-provided taxonomy. The taxonomy (i.e., a set of is-a hierarchies) is used to aggregate low-level data items into higher-level concepts. For example, a city (e.g., Milan) can be generalized as the corresponding country (e.g., Italy). A generalized itemset (e.g., {(Location,Italy),(Day,Working day)}) consists of a set of items that either occur in the source dataset or represent data item generalizations according to the given taxonomy. However, the potentially large set of extracted itemsets could be hardly manageable by domain experts for manual inspection. By comparing itemsets extracted at different abstraction levels we focus the analysis on worthwhile itemset subsets, representing only contrasting or unexpected situations, because in social data analysis targeted actions are often triggered by exceptional or surprising events (Guo et al., 2009; Lin et al., 2009). Specifically, itemsets can be further evaluated and compared by using established correlation indexes (e.g., Kulc, interest, lift Wu et al., 2010), which indicate the item correlation type (i.e., positive, negative, or null) and strength. If the correlation type of an ancestor (high-level) itemset is in contrast with those of many of its low-level descendant itemsets, then an anomalous situation may come out. For example, if items in {(Location,Italy),(Day,Working day)} are positively correlated with each other, we expect that itemsets such as {(Location,Milan),(Day,Working day)} have the same correlation type. Otherwise, the comparison between the two itemsets can be worth investigating more in detail. Even though itemsets with contrasting correlation have already been studied in (Barsky et al., 2011), to the best of our knowledge an approach to mining *large groups* of itemsets in contrast with a common ancestor in terms of correlation type has never been proposed so far.

This paper presents the TFC Analyzer (Twitter Flipping Correlation Analyzer) system to support different and interesting targeted analysis, i.e., topic trend analysis, context-aware service profiling, outlier detection. It aims at analyzing Twitter posts to discover subsets of frequent generalized itemsets that potentially represent contrasting situations. Given the Twitter posts (i.e., the tweets) enriched with their publication context (i.e., date, time, place), a novel kind of pattern, namely the Strong Flipping Generalized Itemset (SFGI), is mined. SFGIs are patterns in the form $X \sim \Psi$, where X is a frequent generalized itemset having a *large* set Ψ of low-level exceptions, i.e., frequent descendant itemsets of X whose correlation type changes with respect to X. The existence of a large group of contrasting low-level itemsets may indicate the presence of an unexpected situation in Twitter data. To extract all SFGIs whose number of contrasting low-level correlations is equal to or exceeds a given (analyst-provided) threshold, TFC Analyzer exploits an efficient LCM-based (Linear Time Closed Itemset Miner-based) itemset mining algorithm combined with an ad-hoc post-pruning phase.

Experiments performed on real-life data coming from Twitter demonstrate the effectiveness of the proposed system in discovering interesting knowledge. Furthermore, the performance and scalability of the adopted mining strategy have been evaluated on real and synthetic datasets.

Even though this work focuses on Twitter data analysis, it is worth mentioning that the proposed patterns can be successfully mined and exploited to support knowledge discovery from data coming from different contexts.

This paper is organized as follows. Section 2 presents a motivating example. Section 3 compares our work with related approaches. Section 4 thoroughly describes the characteristics of the TFC Analyzer system, while Section 5 describes the experiments performed. Section 6 discusses the extension of TFC Analyzer in a distributed environment. Finally, Section 7 draws conclusions and discusses future work.

Table 1
Example dataset $\mathcal{D}$.

Location	Day
Milan	Working day
Milan	Working day
Turin	Working day
Trento	Working day
Naples	High day

2. Motivating example

We are interesting in analyzing Twitter data to efficiently support business applications based on social network data mining, e.g., context-aware service profiling. Let us consider the example Twitter dataset $\mathcal{D}$ in Table 1. It consists of 5 Twitter posts. For each post the publication weekday and the city of provenance of the author are available. For the sake of simplicity, in this preliminary example we disregard the textual content of the tweet as well as any other contextual information.

Fig. 1 shows an example of taxonomy built on the analyzed data. It generalizes cities as the corresponding region and country, whereas publication days (working or high days) are aggregated into *weekday*. Frequent generalized itemsets are sets of items or generalized items that (i) represent high-level correlations among data and (ii) frequently occur in the dataset, i.e., their support value is above a given threshold (Srikant and Agrawal, 1995). Itemset quality measures (e.g., lift, Kulc, interest Wu et al., 2010) have been proposed to evaluate item correlation strength and type (i.e., positive, negative, or null). Since a potentially large number of itemsets can be extracted, a manual inspection of the result set could be a challenging task. To discover unexpected and potentially interesting situations we focus on subsets of item correlations at different abstraction levels that are in contrast in terms of correlation type.

Table 2 reports the frequent generalized itemsets mined from $\mathcal{D}$ by a traditional approach (Srikant and Agrawal, 1995) (see Column 2). The mining task was accomplished by exploiting the taxonomy in Fig. 1 and by enforcing an absolute minimum support threshold min_sup equal to 1. For each itemset, the support value (Column 3), the generalization level (Column 4), and the value of an established correlation index, i.e., the Kulczynsky index (Wu et al., 2010) (hereafter denoted as Kulc), are also reported. By setting the maximum negative and minimum positive Kulc thresholds max_neg_cor and min_pos_cor to 0.7 and 0.8, respectively, itemsets with Kulc between 0.7 and 0.8 are uncorrelated, itemsets with Kulc below 0.7 show negative item correlation, whereas itemsets with Kulc above 0.8 indicate a positive item correlation, i.e., their items co-occur more than expected.

From the comparison between each frequent generalized itemset in Table 2 and its frequent descendants it appears that 3 out of 17 frequent generalized itemsets have at least one exception (i.e., a low-level descendant itemset with different correlation type). The existence of a large number of exceptions may prompt experts to manually explore these pattern subsets, because items unexpectedly change their correlation type while climbing up or down the taxonomy. A SFGI $X \sim \Psi$ combines a frequent generalized itemset X with the corresponding subset Ψ of frequent low-level exceptions. SFGIs mined by setting the minimum number of exceptions min_except to 1 are marked with X at Column 5 in Table 2. Focusing on the generalized itemsets with at least one exception, itemsets (11), (12), and (13) are not yet considered. Let us consider the positively correlated itemset (15) {North Italy, Week Day} and its corresponding negatively correlated descendant itemsets {Turin,